DR-Minerva: a Multimodal Language Model based on Minerva for Diagnostic Information Retrieval

Irene Siragusa[1,2][0009−0005−8434−8729], Salvatore Contino[1][0000−0002−7476−1545], and Roberto Pirrone[1][0000−0001−9453−510X]

[1] Department of Engineering, University of Palermo, Palermo, 90128, Sicily, Italy
[2] Department of Computer Science, IT University of Copenhagen, København S, 2300, Denmark
{irene.siragusa02,salvatore.contino01,roberto.pirrone}@unipa.it

Abstract. This paper illustrates the development of Minerva Diagnostic Retriever (DR-Minerva), a Visual Language Model specialized in the medical domain. Prompted using a textual input with the patient's information along with a CT or MR scan, the model provides information about the body part and the scanning modality of the given image. The model relies on the Flamingo architecture, which is well known for its good in-context and few-shot learning capabilities, and it encodes textual data using Minerva, a novel Large Language Model trained on English and Italian data. Model performances are improved via fine-tuning the aforementioned model, and using external knowledge by means of a Retrieval Augmented Generation approach. At inference time, the model is injected with the retrieved examples in form of in-context learning. The authors developed a rearranged version of the MedPix® multi-modal medical dataset, that was used for both the development and the test of the model as long as for retrieval. A detailed description of the system is reported along with the experimental results that are discussed in thoroughly. Dataset and models used are available on GitHub[3].

Keywords: Multimodal Language Model · Retrieval Augmented Generation · Information Retrieval · Minerva · Flamingo · MedPix

1 Introduction

The spread of Artificial Intelligence (AI) in the medical domain has been revolutionary, beginning to transform the way in which diagnosis, treatment, and monitoring of patients are carried out. In particular, in recent years, the development of AI-based technologies for decision making support for physicians, gained a relevant interest in the scientific community, through the use of increasingly complex and precise Deep Neural Networks (DNNs) capable of analysing the whole variety of data available from clinical departments [5]. Today, the development of increasingly precise support systems is the natural development for

[3] https://github.com/CHILab1/MedPix-2.0

computer applications within this domain, which must, however, not only develop predictive capabilities, but also an ever-increasing level of trustworthiness for physician and patient safety. To achieve these goals, DNNs need to be trained on an ever-increasing amount of data to improve their generalization capability. Unfortunately, the shortage of data is the biggest obstacle that does not allow rapid progress towards this goal to date. Currently, the models developed by the scientific community are based on public domain datasets, which are of poor quality because they are collected episodically without an established protocol for adding new data to the set. Often, such datasets are assembled for a scientific challenge, and their metadata only reflect the purpose of the scientific question behind the competition. The data needed to build reliable AI-enabled Medical Decision Support Systems (MDSS) must be collected directly from clinical sources, and their metadata must be standardized, especially for Vision Language Models (VLM) or on a Multimodal architecture of various kinds (Large Multimodal Model, LMM). It is precisely these models that are perfectly suited to the role required for physician support, thanks to their ability to integrate and process textual and visual information, providing rapid and objective support by analysing the features extracted from the data provided.

We present a new implementation of a VLM, based on Flamingo[2] and Minerva LLM, called Minerva Diagnostic Retriever (DR-Minerva), which have been trained through fine-tuning for the classification and prediction of two main features of biomedical data, i.e. the *modality* that contains the information inherent to the type of source of the biomedical image (e.g. CT, MRI) and the *location* that instead refers to the anatomical region that has been examined. The proposed neural architecture aims to perform few-shot predictions by identifying both modality and location, and if required returns the join of both predictions given a medical image and a short text with the information of the patient. Few-shot prediction is done via a Retrieval Augmented Generation (RAG) approach [15], leveraging on the peculiar textual information provided and the re-arranged version of the MedPix® [1] was used.

The paper is arranged as follows: Section 2 illustrates the relevant contributions within LMM in the medical domain. The architecture of the developed systems is reported in Section 3 along with the dataset used, while the experimental setups and results are reported and discussed in Section 4. Future works and concluding remarks are drawn in Section 5 and 6 respectively.

2 Related works

Since development of transformed-based Language Models (LM) [28] like BERT [6], considerable improvements were done in building LMs till to Large Language Models (LLM). Those systems reach competitive performances with State Of The Art (SOTA) BERT-based models in most of the traditional Natural Language Processing (NLP) tasks [10], but they are intrinsically full of issues. Recent LLMs are very large, averaging from 70B to 175B as for Llama models [27,20] and GPT-3.5 [4], and a full fine-tuning of these models is actually impracticable

due to the high computational cost and resources required. Moreover, there is no open information about the training procedure or the data involved of both models. Despite those issues, the interest of the scientific community is looking towards Large Multimodal Model (LMM), which leverage both textual, visual or audio data. Since Medical Imaging is intrinsically a multimodal domain, and it deeply focuses on analyzing both images and the related reports [8], Visual Language Models are the most used within these applications.

Most of the medical VLM rely on SOTA models, such as CLIP [24], LLaVA [19], and OpenFlamingo [3], an open-source version of Flamingo [2] and via a fine-tuning procedure, they succeed in developing their corresponding medical versions, BiomedCLIP [29], LLaVa-Med [16], and Med-Flamingo [21]. These models share a common pipeline, where images and text are encoded separately with their respectively visual and textual models, and then merged together to generate the textual output.

CLIP is trained to learn a multi-modal embedding space by jointly training both an image encoder and a text encoder to maximize the cosine similarity of the paired image and text embeddings while minimizing the cosine similarity of the incorrect pairings [24].

Both LLaVA and Flamingo rely on the pre-trained CLIP visual encoder to extract the visual features from a given image. LLaVA projects the obtained visual features to the word embedding space via a linear layer to pass them to the LM [19]. Analogously in Flamingo, visual tokens are extracted from the visual features through a "Perceiver Resampler" and then are incorporated with the textual encoding via a cross-attention layer, which is interleaved between the frozen pre-trained LM layers [2].

BiomedCLIP [29], LLaVa-Med [16], and Med-Flamingo, follow the training strategies of the models they are based on, and they are fine-tuned on data sets containing pairs of medical images and their caption, such as MTB [21], PMC-OA [17] and then are evaluated with the medical dataset for Visual Question Answering (VQA) like VQA-RAD [14], PathVQA [9] and SLAKE [18]. Both VQA-RAD and SLAKE collect radiologic images, while PathVQA contains pathology images, and the models are expected to reply to the given question based on the information derived from the proposed image.

3 System description

In the following sub-sections a detailed description of the proposed system is given along with the design choices, and the overview of the dataset is provided along with the metrics used for the evaluation phase.

3.1 RAG-based Flamingo

The overall DR-Minerva architecture is shown in Fig. 1. The system relies on both the Flamingo architecture [2] and Minerva [23], a novel LLM trained from scratch on English and Italian data as part of the activities in the PNRR FAIR

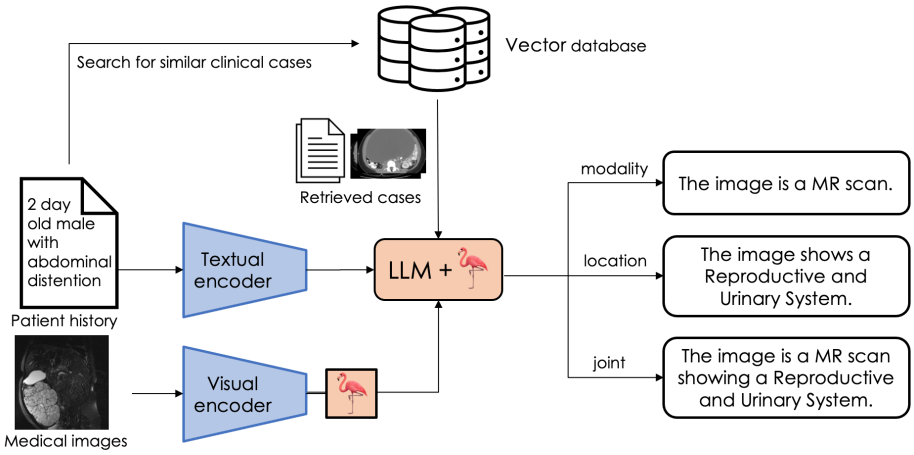

Fig. 1. Overview of DR-Minerva architecture.

Transversal Project 2: "Vision, Language and Multimodal Challenges"[4]. The project is an effort made by almost twenty Italian Universities, and the authors are involved in the research for developing LMMs tailored for specific domains. We chose Minerva since it is a completely open-source model, since training set, architecture and weights are freely available, and its training set is half in English and half Italian, making it suitable for further experiments with the developed architecture and Italian data.

We used Open Flamingo [3], the open-source version of Flamingo, in the 3B parameters version[5]. As it is well known, the Flamingo architecture exhibits good in-context learning capabilities that make it suitable to adapt in diverse domains [2]. We maintained the CLIP ViT-L/14 [24] as the visual encoder, and we adopted Minerva-3B[6] as language encoder. The overall model is queried using the prompt reported in Table 1 to instruct the model on how to behave and generate the desired output.

We developed a suitable RAG component for our system [15] so, at inference time, the model is provided with an enriched prompt that can improve its performance. Since the AI models are required to be as precise as possible, in particular in medical domain, we query DR-Minerva with both the target medical image and a template built from some personal information of the patient (e.g. age and sex) followed by the history of the patient in order to prevent empty textual samples. Then the closest clinical cases are retrieved w.r.t. the patient's history, and they are attached to the prompt as few-shot learning examples.

[4] https://fondazione-fair.it/en/transversal-projects/tp2-vision-language-and-multimodal-challenges/
[5] https://huggingface.co/openflamingo/OpenFlamingo-3B-vitl-mpt1b
[6] https://huggingface.co/sapienzanlp/Minerva-3B-base-v1.0

The RAG module picks (up to) the four closest clinical cases to the query, and it is build via the LangChain framework[7] that in turn uses a FAISS [7] vector database where data are stored using Linq-Embed-Mistral [13]. This is recognized as the best model in Massive Text Embedding Benchmark (MTEB) [22] for Information Retrieval[8].

To effectively evaluate RAG performance, we run the evaluation experiments with and without adding the retrieved examples, at inference time. It is worth noticing that in the no-RAG configuration, two examples showing two different scanning modalities are provided to the model to guide it at generation phase. During the developing process, some experiments were done querying the model providing just the instruction, i.e. in zero-shot configuration, as it is reported in section 4.

Table 1. The structure of the prompt is reported as well as the template of the corresponding expected answers

Type	Prompt	Response
Modality	"Given the following medical images and the patient history, provide information about the scanning modality."	The image is a MR scan.
Location	"Given the following medical images and the patient history, provide information about the body part shown in the image."	The image shows a head.
Join	"Given the following medical images and the patient history, provide information about the scanning modality and the body part shown in the image."	The image is a MR scan showing a head.

3.2 The Dataset

To develop DR-Minerva, we used a re-arranged version of the MedPix® dataset [26] [9]. MedPix® [1] is a multimodal semi-structured dataset of clinical cases released by the National Institutes of Health (NIH). For each case, a clinical report of the patient is reported along with some generic information about the disease, and some medical images with additive information related to both the scanning modality and the body part.

MedPix® dataset is freely available but the textual information is not provided in a suitable format for training AI system. The re-arranged version we used, collects all the information and structure them in two JSON files, namely the case-topic and the description file. The former collects the clinical cases, identified by their *uid* code, all the patient information, such as age, sex and her/his history, the diagnostic finding and the suggested treatment (case information), and general information about the disease from an academic point of view (topic information). The latter collects the caption of the medical images: for each image the *uid* code is reported, along with the scanning modality, the caption of the image and the body part shown. The aforementioned information

[7] `https://www.langchain.com/`

[8] as in `https://huggingface.co/spaces/mteb/leaderboard` in June 2024

[9] `https://github.com/CHILab1/MedPix-2.0`

are reported as key-value pair following the JSON format: values are textual strings that can be a single word, as for the scanning modality, or a paragraph as for the differential diagnosis in the case information.

Given the data in a more accessible format, it was possible to select the relevant pieces of information and consequently create the classification tasks. We decided to focus on three evaluation setting, namely *modality*, *location* and *join*. The scanning modality evaluation consist in determining if a given image is a Computed Tomography (CT) or a Magnetic Resonance (MR) scan, while the location requires to individuate the body part shown, and the join setting asks for information about both modality and location. A single location is assigned to each image and refers to the macro-area shown[10] and are Thorax, Head, Abdomen, Reproductive and Urinary System and Spine and Muscles. The original dataset considers also General and Nervous System labels, but, since samples belonging to these categories were few compared to the others, images belonging to these location were re-assigned by a specialist to the aforementioned five. There is also worth mentioning that each image is associated to one location and one scanning modality, while a clinical case can contain multiple images, obtained with different scanning modality and showing different body part.

Thus, for each clinical case, we considered the history of the patient, his/her sex and age, and we created a sample document that, within the annotated images with modality and location labels, constitute the multimodal dataset for DR-Minerva. A sample of the dataset is reported in the figure below.

Example of some of the structured information from MedPix 2.0

UID: *MPX1463*
CT: *MPX1463_synpic17575*
MRI: *MPX1463_synpic17576, MPX1463_synpic17577*
Age: *6 y.o.*
Sex: *female*
History: *A 6 YO female presented to the emergency department with a 3 day history of right eye pain and severe decreasing visual acuity.*
Disease: *Optic Neuritis*
Modality: *MR*
Location: *Head*

RAG document

6 y.o. female patient suffering from a Optic Neuritis (imaging and clinical findings combined to verify diagnosis) .
A 6 YO female presented to the emergency department with a 3 day history of right eye pain and severe decreasing visual acuity.

Inference query

6 y.o. female patient. A 6 YO female presented to the emergency department with a 3 day history of right eye pain and severe decreasing visual acuity.

Fig. 2. A simple representation of a multimodal sample of the dataset. The actual structure of the RAG documents and the inference query, derived from the JSON file of the dataset are reported.

[10] a more detailed location information is provided but not considered

A semi-automatic 80%-10%-10% split was built from the documents keeping images from the same clinical case in the same split and assuring a balance for modality and location labels, as it is shown in Table 2. Maintaining the balance, we further split the training set in train-1 and train-2. Demographics and history of the patient and the clinical diagnosis from train-1, following ad hoc designed template, are saved as documents (e.g. .txt file) and used to build the vector database and used as retriever corpus at inference time; train-2 samples are used for Flamingo fine-tuning, while both training splits are used for Minerva fine-tuning. At inference time, the model is queried to provide information about the *modality* the image is captured, the body part *location* shown and both characteristics (*join*). At inference time, depending on the evaluation setting, the designed prompt reported in Table 1 is used, along with the retrieved documents and the corresponding images, from train-1 split, the inference image and the textual input, constructed with a template reporting sex, age and history of the patient, as in Figure 1.

To the best of our knowledge, there is no available dataset and related classification tasks, that covers such a variety of information regarding scanning modality and body part. The majority of dataset focuses on a scanning modality or body part, like chest [12] or brain [11], thus making impossible to develop more complex tasks where diverse scanning modality and body part are jointly taking into account and that shares also homogeneous textual information that can be used for train Multimodal Models.

Table 2. Below an overview of the used dataset is provided. We refer to Reproductive and Urinary System as RaUS and to Spine and Muscles as SaM (inside the brackets the number of images is reported).

Train	*Dev*	*Test*
• Images (1653)	• Images (197)	• Images (200)
∗ TAC (878)	∗ TAC (84)	∗ TAC (100)
∗ MRI (775)	∗ MRI (113)	∗ MRI (100)
• Location	• Location	• Location
∗ Thorax (263)	∗ Thorax (30)	∗ Thorax (41)
∗ Head (742)	∗ Head (66)	∗ Head (76)
∗ Abdomen (264)	∗ Abdomen (23)	∗ Abdomen (32)
∗ RaUS (127)	∗ RaUS (20)	∗ RaUS (11)
∗ SaM (257)	∗ SaM (58)	∗ SaM (40)

4 Experiments

4.1 Experimental set-up

The whole architecture was developed on server with 96 Intel(R) Xeon(R) Gold 6442Y CPUs and 2 48 GB NVIDIA RTX 6000 Ada Generation.

We instruction-tuned Minerva for 20 epochs on a single GPU taking approximately 4 hours: we follow Alpaca-LoRA[11] setting and trained the model with all the training set. Starting from the multimodal split, we created the training sample by adding to the instruction in Table 1, the patient demographics and history. We further added the Flamingo's special tokens `<image>` and `<|endofchunk|>` generating a training set of 4959 samples, three times larger than the original since samples where created for the three query, namely modality, location and join.

Due to computational restrictions, we fine-tuned Flamingo, with the fine-tuned version of Minerva, per 10 epochs over the CPU with batch size of 1, following the train hyperparameters of Open Flamingo[12]. The whole process last approximately 3 days.

As for inference, runs where no RAG context was considered took 20 minutes on average on a single GPU, while the ones with RAG, approximately one hour and half and samples with a bigger RAG context that didn't fit the GPU, where manually queried in CPU.

4.2 Metrics

The created evaluation tasks can be considered as multi-class classification tasks, which performances can be evaluated with classical classification metrics such as accuracy, precision, recall and F1 score, after an output standardization phase. Since the predicted labels came out from a generative model, it is necessary to check if the generated output matches with or contains one of the possible labels: if an exact match is found, a valid predicted label can be associated with the generated text, otherwise an error label is assigned. This assignment is necessary for metrics calculation and it provides the information that the model generates something unmeaning, like a label spelled incorrectly, not generated at all or invalid, that is the case where a plausible label is generated, but does not meet the task constrains. E.g. at inference time for modality evaluation, if the model labels a image as a PET, it would be considered as an error since the task considers only CT and MR as possible scanning modalities. For precision, recall and F1 score, macro average is considered.

4.3 Results

In the subsections below are reported the obtained experimental results over the test set and a relative discussion for each evaluation setting. As for the evaluation with the fine-tuned version of Flamingo, we report the evaluation after 5 epochs (flamingo-ft-5) and after 10 epochs (flamingo-ft-10). Regardless the evaluation setting, at least a one-shot example should be provided to the model to generate a meaningful answer. Experiments with only prompt and query, not only are unsatisfactory, but also lack of consistency since the model start

[11] `https://github.com/tloen/alpaca-lora`
[12] `https://github.com/mlfoundations/open_flamingo`

generating unmeaning output and it is not following the provided instruction. In general, 5 epochs of fine-tuning Flamingo are enough for reaching satisfactory results and further training can decrease its performances.

Table 3. Experimental results for test set in modality evaluation setting. The starred line represent the inference in a zero-shot evaluation mode.

LLM model	**Flamingo version**	**RAG**	**Accuracy**	**Precision**	**Recall**	**F1 score**
Minerva-3B *	flamingo-base *	*	0.17 *	0.174 *	0.113 *	0.137*
Minerva-3B	flamingo-base		0.36	0.369	0.24	0.278
Minerva-3B	flamingo-base	x	0.305	0.419	0.203	0.194
Minerva-3B ft	flamingo-base		0.405	0.463	0.27	0.341
Minerva-3B ft	flamingo-base	x	0.425	0.322	0.283	0.264
Minerva-3B ft	flamingo-ft-5		0.865	0.894	0.865	0.862
Minerva-3B ft	flamingo-ft-5	x	0.88	0.593	0.587	0.59
Minerva-3B ft	**flamingo-ft-10**		**0.935**	**0.94**	**0.935**	**0.935**
Minerva-3B ft	flamingo-ft-10	x	0.82	0.583	0.547	0.551

Modality evaluation setting In table 3 are collected the experimental results over the test set for modality evaluation setting. In this evaluation setting, no significant improvement is provided by the RAG module, while performances highly benefit from Flamingo fine-tuning and the model fine-tuned over 10 epochs without RAG, reaches the best performances for every considered metric, reaching at least 0.93 for accuracy, precision, recall and F1 score. Overall these performances are not surprising, since this evaluation setting is the easiest one and can be considered as a binary classification task as it is structured: generally speaking, CT and MR are not the only possibilities as for scanning modality, but in the used dataset those are the only two option considered.

In order to verify the effectiveness of using the visual encoder of Flamingo, a comparison with state-of-the-art approaches was conducted. In fact, in the task modality the results obtained with Flamingo are higher than those reported at the state of the art in the paper by Raffy et al. [25] The latter in the diagnostic modality classification obtained an average Recall value of 92.4% (92.5% for CT and of 92.3% for MRI) lower than the performance achieved by our model.

Location evaluation setting In table 4 are collected the experimental results over the test set for location evaluation setting. Given the starting accuracy of 0.03 of the base model with RAG, the best results are obtained with Flamingo fine-tuned over 5 epochs, reaching an accuracy of 0.72 while precision, recall and F1 are stable around 0.51. Differently from the modality evaluation setting, this is a harder task for the model that, despite enhancing its performances via the usage of RAG, cannot reach highly satisfactory results, compared to the ones in Table 3, but surprisingly high compared to the starting point.

In this case further fine-tuning of Flamingo for 10 epochs, is not beneficial for the model, by contrast we aim that the textual information provided along with the image for location classification, improves and guide the general-purpose visual encoder that do not use any segmentation techniques to analyze the provided images for sub-sequentially classification.

Table 4. Experimental results for test set in location evaluation setting. The starred line represent the inference in a zero-shot evaluation mode.

LLM model	Flamingo version	RAG	Accuracy	Precision	Recall	F1 score
Minerva-3B*	flamingo-base*	*	0.0*	0.0*	0.0*	0.0*
Minerva-3B	flamingo-base		0.0	0.0	0.0	0.0
Minerva-3B	flamingo-base	x	0.03	0.292	0.022	0.041
Minerva-3B ft	flamingo-base		0.0	0.0	0.0	0.0
Minerva-3B ft	flamingo-base	x	0.035	0.146	0.015	0.028
Minerva-3B ft	flamingo-ft-5		0.285	0.245	0.242	0.142
Minerva-3B ft	**flamingo-ft-5**	**x**	**0.72**	**0.524**	**0.507**	**0.512**
Minerva-3B ft	flamingo-ft-10		0.39	0.365	0.272	0.225
Minerva-3B ft	flamingo-ft-10	x	0.625	0.479	0.464	0.457

Join evaluation setting In table 5 are collected the experimental results over the test set for join evaluation setting, that is considered yet for the combined prediction task yet for the separate evaluations of both modality and location. Here can be clearly be notice how the combined prediction of modality and location and location alone can be difficult for the model, and how can be substantially improved with the RAG module, as for ACC-J that grows from 6.50% to 25.50% and as for ACC-L from 7.50% to 55.00% without any fine-tuning. As expected, fine-tuning Flamingo leads to the best performances that, as for the join evaluation setting, that reaches an accuracy of only 0.65: this result, shows that there is a substantial room for improvement in this evaluation setting and it confirms that, the multi-label classification task for determining the location of a given images, i.e. the location evaluation, is really challenging for the model despite five macro body part are considered.

5 Future works

The final objective of our work is to build an AI system that can effectively help physicians during the diagnostic process, not only providing a detailed explanation of the proposed clinical image, but also adding some general information, about the found lesion or the disease. We are currently working on a second version of the model that, in a purely generative setting, provides a free-text clinical report, leveraging a re-defined RAG. The idea is to retrieve not only clinical reports, but also general information about the diseases, coming from

Table 5. Experimental results for test set in join evaluation setting, EM stands for exact match and ACC stands for accuracy, PREC for precision and REC for recall. ACC-M, ACC-L and ACC-J refers to the accuracy with the **M**odality, **L**ocation and **J**oin evaluation setting respectively. The starred line represent the inference in a zero-shot evaluation mode.

LLM model	Flamingo version	RAG	ACC-M	ACC-L	ACC-J	PREC-M	PREC-L	PREC-J	REC-M	REC-L	REC-J	F1-M	F1-L	F1-J
Minerva-3B*	flamingo-base*	*	0.25*	0.0*	0.0*	0.194*	0.0*	0.0*	0.167*	0.0*	0.0*	0.179*	0.0*	0.0*
Minerva-3B	flamingo-base		0.2	0.075	0.065	0.129	0.0255	0.0121	0.133	0.061	0.0303	0.131	0.036	0.0173
Minerva-3B	flamingo-base	x	0.45	0.55	0.255	0.168	0.517	0.173	0.3	0.396	0.19	0.215	0.426	0.163
Minerva-3B-ft	flamingo-base		0.5	0.265	0.19	0.412	0.201	0.0631	0.333	0.176	0.183	0.298	0.127	0.0769
Minerva-3B-ft	flamingo-base	x	0.51	0.625	0.3	0.417	0.46	0.277	0.34	0.437	0.2	0.284	0.42	0.166
Minerva-3B-ft	**flamingo-ft-5**		0.925	0.425	0.37	0.935	0.621	**0.462**	0.925	0.365	0.304	0.925	0.338	0.235
Minerva-3B-ft	**flamingo-ft-5**	x	0.89	**0.72**	**0.65**	0.597	**0.622**	0.434	0.593	**0.632**	**0.389**	0.595	**0.613**	**0.382**
Minerva-3B-ft	**flamingo-ft-10**		**0.93**	0.55	0.5	**0.936**	0.512	0.401	**0.93**	0.382	0.366	**0.93**	0.362	0.302
Minerva-3B-ft	flamingo-ft-10	x	0.845	0.675	0.565	0.588	0.511	0.37	0.563	0.504	0.37	0.564	0.488	0.34

the same dataset or built from high-quality sources as textbooks, and arranged in a graph database, thus resembling a medical Knowledge Base that can be better navigated to reach the relevant documents for report generation. To develop this model, we will use Leonardo supercomputer[13] via a ISCRA-C application.

Another objective relies on the choice of the LLM used: Minerva is trained from scratch from English and Italian data, and our purpose is to develop and analyze an Italian version of DR-Minerva via a translation procedure of the dataset used, i.e. and Italian version of MedPix 2.0.

6 Conclusions

We presented DR-Minerva, a Multimodal Language Model for medical domain based on Minerva LLM that employs a RAG based approach to enhance its classification capabilities for classifying a medical image considering the scanning modality, the body part shown or both. Our experiments reveal that the join proposed task for determine both modality and location is challenging for the model and there is room for improvement with the given task setting and for further generation-related tasks in free-text modality.

Acknowledgments. We would like to thank the Sapienza NLP group for developing Minerva LLM. This work is supported by the cup project B73C22000810001, project code ECS_00000022 "SAMOTHRACE" (Sicilian MicronanoTech Research And Innovation Center).

Disclosure of Interests. The authors declares that they have no relevant or material financial interests that relate to the research described in this paper.

References

1. Medpix, `https://medpix.nlm.nih.gov/home`

[13] `https://leonardo-supercomputer.cineca.eu/it/home-it/`

2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022)
3. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
5. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. Future Healthcare Journal **6**(2), 94–98 (Jun 2019). https://doi.org/10.7861/futurehosp.6-2-94
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024), https://arxiv.org/abs/2401.08281
8. Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. arXiv preprint arXiv:2403.02469 (2024)
9. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
10. Hromei, C.D., Croce, D., Basile, V., Basili, R.: Extremita at evalita 2023: Multi-task sustainable scaling to large language models at its extreme. In: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023). CEUR.org, Parma, Italy (September 2023)
11. Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L. Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
12. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
13. Junseong, K., Seolhwa, L., Jihoon, K., Sangmo, G., Yejin, K., Minkyung, C., Jyyong, S., Chanyeol, C.: Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog (2024), https://getlinq.com/blog/linq-embed-mistral/
14. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
15. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for

knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
16. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
17. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
18. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654. IEEE (2021)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
20. Llama Team, A..M.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
21. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023)
22. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316 (2022). https://doi.org/10.48550/ARXIV.2210.07316, https://arxiv.org/abs/2210.07316
23. Orlando, R., Moroni, L., Huguet Cabot, P.L., Conia, S., Barba, E., Navigli, R.: Minerva technical report (2024), https://nlp.uniroma1.it/minerva/
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
25. Raffy, P., Pambrun, J.F., Kumar, A., Dubois, D., Patti, J.W., Cairns, R.A., Young, R.: Deep learning body region classification of mri and ct examinations. Journal of Digital Imaging **36**(4), 1291–1301 (Aug 2023). https://doi.org/10.1007/s10278-022-00767-9
26. Siragusa, I., Contino, S., Ciura, M.L., Alicata, R., Pirrone, R.: Medpix 2.0: A comprehensive multimodal biomedical dataset for advanced ai applications. In: Proceedings of the 3rd Italian Conference on Big Data and Data Science, ITADATA2024 (to appear). Pisa, Italy (September 2024), https://arxiv.org/abs/2407.02994
27. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)

28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
29. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)