

Bootstrapping Large Language Models for Radiology Report Generation

Chang Liu¹, Yuanhe Tian², Weidong Chen¹, Yan Song^{1*}, Yongdong Zhang¹

¹University of Science and Technology of China

²University of Washington

lc980413@mail.ustc.edu.cn, yhtian@uw.edu, chenweidong@ustc.edu.cn,
clksong@gmail.com, zhyd73@ustc.edu.cn

Abstract

Radiology report generation (RRG) aims to automatically generate a free-text description from a specific clinical radiograph, e.g., chest X-Ray images. Existing approaches tend to perform RRG with specific models trained on the public yet limited data from scratch, where they often lead to inferior performance owing to the problem of inefficient capabilities in both aligning visual and textual features and generating informative reports accordingly. Currently, large language models (LLMs) offered a promising solution to text generation with their power in learning from big data, especially for cross-modal scenarios such as RRG. However, most existing LLMs are pre-trained on general data, and suffer from the same problem of conventional approaches caused by knowledge gap between general and medical domain if they are applied to RRG. Therefore in this paper, we propose an approach to bootstrapping LLMs for RRG with an in-domain instance induction and a coarse-to-fine decoding process. Specifically, the in-domain instance induction process learns to align the LLM to radiology reports from general texts through contrastive learning. The coarse-to-fine decoding performs a text elevating process for those reports from the ranker, further enhanced with visual features and refinement prompts. Experimental results on two prevailing RRG datasets, namely, IU X-Ray and MIMIC-CXR, demonstrate the superiority of our approach to previous state-of-the-art solutions. Further analyses illustrate that, for the LLM, the induction process enables it to better align with the medical domain and the coarse-to-fine generation allows it to conduct more precise text generation.

1 Introduction

Medical imaging plays an important role in clinical diagnosis and treatment recommendation, where physicians normally write reports according to the syndromes of patients that are reflected in images so as to form professional records. As a special type of medical images, radiographs are essential in evaluating patients' medical condition with analyzing internal structures of their bodies, and have been widely used in orthopedics, dentistry, cardiology,

and pulmonology, etc. Yet, writing radiology reports is a time-consuming job, and always error-prone for inexperienced radiologists, which motivate a series of studies (Jing, Xie, and Xing 2018; Chen et al. 2020, 2021; Qin and Song 2022; Tanida et al. 2023; Liu, Tian, and Song 2023) on automatic reports generation. They have achieved great success on this topic that has emerged as an attractive research direction in both artificial intelligence and clinical medicine.

Aforementioned approaches for radiology report generation (RRG) employ the encoder-decoder architecture and mainly focus on improving the capabilities of cross-modal alignment and text generation, which are fundamentally restricted when they are learned from a rather small and fixed set of radiograph-report pairs. Consider recent large language models (LLMs) illustrate their superiority in generating high-quality text with few examples, it is expected to apply LLMs in RRG by fine-tuning with limited data.¹ However, in doing so, one faces a challenging barrier of domain variance since ready-to-use LLMs are often pre-trained on general data, which causes a series of problems such as ill-representing visual and textual features from in-domain data, generating texts without domain characteristics, etc. Therefore, domain adaptation and refined text generation are expected on LLMs for in-domain applications, thus requiring particular LLM optimization processes for RRG.

In this paper, we propose an approach to bootstrapping LLM for RRG, with two components designed for domain adaptation and task-specific generation, namely, in-domain instance induction and coarse-to-fine decoding. Specifically, in-domain instance induction adapts the LLM with learning on radiology reports-alike data, equipped with two parts, related instance retrieval and contrastive semantic ranking. Related instance retrieval provides a series of reports with ranked semantic relations to the input radiograph, and these reports are used in contrastive semantic ranking as related instances, comparing with less correlated instances from other medical sources. The LLM in this induction process is then learned to generate reports similar to high-ranked instances than lower ones, so as to fast align with in-domain and task-specific data. The coarse-to-fine decoding process

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

*Corresponding author.

Code and models of our approach are available at <https://github.com/synlp/R2-LLM>.

¹In practice, fine-tuning LLMs for in-domain tasks still requires a rather large amount of labeled data for better performance when LLMs are not pre-trained on the data from particular domains.

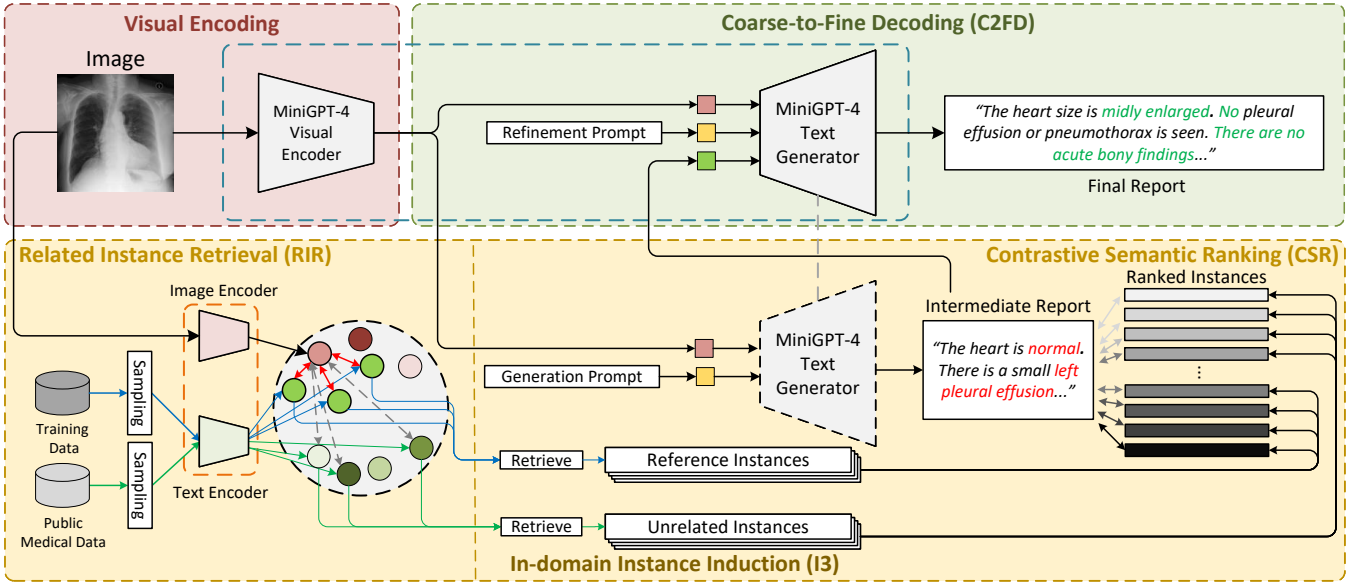


Figure 1: The overall architecture of our proposed approach based on MiniGPT-4 for RRG. The approach consists of three main components, namely, visual encoding, in-domain instance induction (I3), and coarse-to-fine decoding (C2FD), which are represented in red, yellow, and green backgrounds, respectively. The dashed blue box cross visual encoding and C2FD refers to MiniGPT-4. The dashed orange box in the left part of in-domain instance induction stands for the toolkit to retrieve instances from different sources, where the darker colored instances in the output are more related to the input radiograph than the lighter ones. Note that the text generator (LLM) of MiniGPT-4 is shared in I3 and C2FD, indicated by a gray dashed line.

optimizes the LLM to refine intermediate reports generated from the induction process to the precise final reports. Experimental results and analyses on two RRG benchmark datasets, i.e., IU X-RAY (Demner-Fushman et al. 2016) and MIMIC-CXR (Johnson et al. 2019) demonstrate the superiority of our approach, which outperforms strong baselines and achieves state-of-the-art performance on both datasets.

2 The Approach

The overall architecture built upon MiniGPT-4 (Zhu et al. 2023)² is illustrated in Figure 1, with three main components, namely, visual encoding, in-domain instance induction (I3), and coarse-to-fine decoding (C2FD).

The visual encoding process employs an encoder f_{ve} (i.e., MiniGPT-4 visual encoder) to extract latent representations of an input radiograph \mathcal{I} . Then, the in-domain instance induction process uses a text generator f_{tg} (i.e., MiniGPT-4 text generator, namely, Vicuna (Chiang et al. 2023)) to generate intermediate reports based on the resulted radiograph representation from the visual encoder and a generation prompt \mathbf{p}_g , where in training stage, related and unrelated reports are retrieved and sampled to provide a set of ranked instances, for the induction process to update the generator with domain-specific knowledge. Finally, the coarse-to-fine decoding process uses the same text generator f_{tg} to produce

²LLaVA (Liu et al. 2023) is also replaceable and claims similar results in our experiments. We choose MiniGPT-4 because it more focuses on generation tasks that are more relevant to the RRG task, whereas LLaVA is specialized in vision-language understanding tasks according to how they are applied in their original settings.

the final report $\hat{\mathcal{Y}}$ based on the resulted intermediate reports, along with the radiograph representations and a refinement prompt \mathbf{p}_r . Thus, RRG in our approach is formalized as

$$\hat{\mathcal{Y}} = f_{tg}(f_{ve}(\mathcal{I}), \mathbf{p}_r, f_{tg}(f_{ve}(\mathcal{I}), \mathbf{p}_g)) \quad (1)$$

where $\hat{\mathcal{Y}}$ is the final report and \mathcal{I} is the input radiograph. In training, the model is optimized based on the in-domain instance induction loss \mathcal{L}_{I3} and the cross-entropy loss \mathcal{L}_{C2FD} from the generated final reports $\hat{\mathcal{Y}}$ and the gold standard \mathcal{Y}^* , therefore resulting the final loss \mathcal{L} as

$$\mathcal{L} = \beta_1 \mathcal{L}_{I3} + \beta_2 \mathcal{L}_{C2FD} \quad (2)$$

where β_1 and β_2 are hyper-parameters to balance the contribution of the losses. In the following text, we introduce each component according to the aforementioned processing sequence in details, including visual encoding, in-domain instance induction, and coarse-to-fine decoding, respectively.

2.1 Visual Encoding

The MiniGPT-4 visual encoder f_{ve} consists of three modules, namely, the vision transformer f_v (Dosovitskiy et al. 2021), the Q-Former f_q (Li et al. 2023), and a linear projection layer, formulated as $f_{ve}(\cdot) = Linear(f_q(f_v(\cdot)))$. Therefore, radiographs are firstly fed into the vision transformer model f_v , with their features \mathbf{h}_v extracted through

$$\mathbf{h}_v = f_v(\mathcal{I}) \quad (3)$$

and are further processed by the Q-Former f_q to transfer visual representations into textual semantic space by

$$\mathbf{h}_q = f_q(\mathbf{h}_v) \quad (4)$$

where \mathbf{h}_q represents the output features. Finally, a linear projection layer is used to further project \mathbf{h}_q into the latent representations \mathbf{v} to align \mathbf{h}_q to the dimension of hidden states in the text generator through

$$\mathbf{v} = \text{Linear}(\mathbf{h}_q) \quad (5)$$

where \mathbf{v} is used in I3 and C2FD as visual features to guide the generation process of intermediate and final reports.

2.2 In-domain Instance Induction

Once the visual features are extracted, the next step is to generate a corresponding radiology report through a text generator, i.e., an LLM in our approach. In doing so, it still struggles to produce highly patternized radiology reports with professional medical terminology if the generator is trained in the general domain, where misalignment occurs between the general and medical semantic spaces. However, domain adaptation for LLMs is not a trivial task, especially when in-domain training data is limited. Inspired by studies (Ouyang et al. 2022; Bai et al. 2022; Touvron et al. 2023a) that optimize LLMs by learning from ranked texts, we propose a novel method, in-domain instance induction, to bootstrap LLMs with effective domain adaptation based on two sequentially connected components, namely, related instance retrieval and contrastive semantic ranking, whose details are presented in the following texts.

Related Instance Retrieval (RIR) The first component serves as the data provider that offers a series of in-domain references for LLM domain adaptation. In RIR, we retrieve in-domain data from two sources, namely, the training data of RRG and public medical corpora, to provide task-specific and ranking-support references, respectively, where each source contributes M instances ($2M$ instances in total). This two-source design allows the text generator to not only learn domain information by being optimized on in-domain data, but also be equipped with task-specific information by learning to distinguish related instances from others. Specifically, in the retrieval process from the training data, we randomly sample a set of radiology reports $\mathcal{R} = \{R_1, \dots, R_N\}$ with a retrieval size of N . Next, we utilize an image encoder f'_{ve} and its corresponding text encoder f'_{te} from an off-the-shelf toolkit (i.e., MedCLIP (Wang et al. 2022b)) to project the input radiograph \mathcal{I} and the sampled reports \mathcal{R} into the same semantic space through

$$\mathbf{v}_r = f'_{ve}(\mathcal{I}) \quad (6)$$

and

$$\{\mathbf{u}_1, \dots, \mathbf{u}_N\} = f'_{te}(R_1, \dots, R_N) \quad (7)$$

where \mathbf{v}_r and $\{\mathbf{u}_1, \dots, \mathbf{u}_N\}$ are visual and text representations, respectively. Afterwards, for each text representation \mathbf{u}_i , we compute its cosine similarity score c_i with \mathbf{v}_r and select the top M ones based on c_i , then retrieve their corresponding reports from the training data. For the rest M instances from public medical corpora, we randomly sample them out and compute their cosine similarity scores with \mathbf{v}_r following the same process as that we do for c_i . Finally, we merge the two aforementioned instance lists into $\mathcal{S} = [S_1 \dots S_{2M}]$, with the first M instances from public medical corpora, and the rest M ones from the training data.

DATASET	IU X-RAY			MIMIC-CXR		
	TRAIN	VAL	TEST	TRAIN	VAL	TEST
IMAGE	5.2K	0.7K	1.5K	369.0K	3.0K	5.2K
REPORT	2.8K	0.4K	0.8K	222.8K	1.8K	3.3K
PATIENT	2.8K	0.4K	0.8K	64.6K	0.5K	0.3K
AVG. LEN.	37.6	36.8	33.6	53.0	53.1	66.4

Table 1: The statistics of the two benchmark datasets w.r.t. their training, validation, and test sets, including the numbers of images, reports, and patients, and the averaged word-based length (AVG. LEN.) of all reports in each category.

Contrastive Semantic Ranking (CSR) With the retrieved instance list \mathcal{S} , LLM learns to rank them in a contrastive manner by learning over and comparing the representations of LLM output $\hat{\mathcal{Z}}$ and \mathcal{S} . Firstly, by using the visual feature \mathbf{v} of the input radiograph obtained from Eq. (5) and the generation prompt \mathbf{p}_g , we generate the intermediate report $\hat{\mathcal{Z}} = \hat{z}_1 \dots \hat{z}_{N_z}$ with N_z tokens through

$$\hat{\mathcal{Z}} = f_{tg}(\mathbf{v}, \mathbf{p}_g) \quad (8)$$

where the representation \mathbf{o}_n of the n -th token \hat{z}_n is extracted from the last layer³ of f_{tg} by

$$\mathbf{o}_n = f_{tg}(\mathbf{v}, \mathbf{p}_g; \hat{z}_1 \dots \hat{z}_{n-1}) \quad (9)$$

Then, we compute the mean pooling of all \mathbf{o}_n and use the resulting vector \mathbf{o} to represent $\hat{\mathcal{Z}}$. For each instance S_m in \mathcal{S} , we perform a similar process as that we do for $\hat{\mathcal{Z}}$ to obtain its representation \mathbf{o}'_m , with the representation of each token in S_m computed through

$$\mathbf{o}'_{m,n} = f_{tg}(\mathbf{v}, \mathbf{p}_g; s_{m,1} \dots s_{m,n-1}) \quad (10)$$

where $s_{m,1} \dots s_{m,n-1}$ are the $n-1$ tokens prior to the current token as that we do in Eq. (9). Afterwards, to facilitate the learning process in a contrastive manner, we construct M instance pairs $\mathcal{SP} = [(S_1, S_{1+M}), \dots, (S_M, S_{2M})]$, where in each pair, the former comes from public medical corpora and the latter from the training data, which guarantees contrast between unrelated and related instances in each pair. Finally, we compute \mathcal{L}_{I3} through a pairwise optimization

$$\mathcal{L}_{I3} = \frac{1}{M} \sum_{m=1}^M \left[\|\mathbf{o} - \mathbf{o}'_{m+M}\| - \|\mathbf{o} - \mathbf{o}'_m\| + \alpha \right] \quad (11)$$

where $\|\cdot\|$ computes the Euclidean norm of a vector, α is a positive real number that controls the margin. For further explanation, Eq. (11), f_{tg} learns to update the LLM by generating intermediate reports that are closer to the high-ranking instance S_{m+M} in every pair (S_m, S_{m+M}) , so that f_{tg} is gradually aligned to both the medical domain as well as the reports related to the RRG task.

³For the sake of simplicity, in Eq. (9) and (10), we still use f_{tg} to represent representation computation, which is actually different from the generation since the last linear projection layer is omitted.

DATA	MODEL	NLG METRICS							CE METRICS		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	AVG. Δ	P	R	F1
IU X-RAY	MINIGPT-4	0.207	0.068	0.022	0.007	0.091	0.161	-	-	-	-
	+ I3	0.277	0.163	0.122	0.083	0.107	0.181	47.2%	-	-	-
	+ C2FD	0.263	0.154	0.113	0.075	0.101	0.169	28.6%	-	-	-
	+ I3+C2FD (THIS WORK)	0.326*	0.204*	0.154*	0.119*	0.132*	0.234*	57.5%	-	-	-
	MINIGPT-4 (FT)	0.389	0.262	0.181	0.134	0.169	0.308	-	-	-	-
	+ I3	0.458	0.296	0.210	0.158	0.183	0.357	12.8%	-	-	-
	+ C2FD	0.437	0.289	0.205	0.151	0.178	0.336	9.4%	-	-	-
	+ I3+C2FD (THIS WORK)	0.499*	0.323*	0.238*	0.184*	0.208*	0.390*	22.0%	-	-	-
MIMIC -CXR	MINIGPT-4	0.135	0.048	0.014	0.003	0.064	0.104	-	0.126	0.103	0.113
	+ I3	0.212	0.120	0.075	0.031	0.106	0.175	58.2%	0.157	0.113	0.131
	+ C2FD	0.178	0.084	0.051	0.017	0.097	0.163	48.7%	0.144	0.109	0.124
	+ I3+C2FD (THIS WORK)	0.266*	0.152*	0.093*	0.060*	0.109*	0.210*	64.9%	0.228*	0.147*	0.179*
	MINIGPT-4 (FT)	0.323	0.184	0.110	0.069	0.122	0.220	-	0.335	0.246	0.284
	+ I3	0.358	0.203	0.137	0.091	0.139	0.256	14.9%	0.383	0.279	0.323
	+ C2FD	0.345	0.191	0.128	0.079	0.134	0.243	9.2%	0.354	0.251	0.294
	+ I3+C2FD (THIS WORK)	0.402*	0.262*	0.180*	0.128*	0.175*	0.291*	31.5%	0.465*	0.482*	0.473*

Table 2: NLG and CE evaluations of different models on the test sets of IU X-RAY and MIMIC-CXR datasets. “MiniGPT-4 (FT)” denotes the model with MiniGPT-4 pre-trained on MIMIC-CXR training data. “BL” is the abbreviation of BLEU; “MTR” and “RG-L” denote METEOR and ROUGE-L, respectively. The average improvements over all NLG metrics compared to MiniGPT-4 and MiniGPT-4 (FT) correspondingly are also presented in the “AVG. Δ ” column, respectively. “*” marks the results where the improvements are statistically significant over all baselines at $p \leq 0.05$ level.

2.3 Coarse-to-Fine Decoding

Although I3 offers a rather strong learning process to adapt LLM to the medical domain, it is still limited for the LLM to generate precise reports for the RRG task, since only coarse image-text alignment and low-level text references are provided. To further facilitate the generation ability of LLM for RRG, we propose to enhance the LLM with a coarse-to-fine decoding process. Specifically, the text generator⁴ (i.e., the LLM, which is the same as the one in I3) takes the visual representation \mathbf{v} , the refinement prompt \mathbf{p}_r , and the intermediate report $\hat{\mathcal{Z}}$ to generate the final report $\hat{\mathcal{Y}}$ by

$$\hat{\mathcal{Y}} = f_{tg}(\mathbf{v}, \mathbf{p}_r; \hat{\mathcal{Z}}) \quad (12)$$

Afterwards, we compute the cross-entropy loss \mathcal{L}_{C2FD} based on the generated and the gold standard reports for each input radiograph and combine it with \mathcal{L}_{I3} based on Eq. (2) to jointly optimize the LLM accordingly during training.

3 Experiment Settings

3.1 Datasets

We conduct our experiments on two conventional benchmark datasets, i.e., IU X-RAY (Demner-Fushman et al. 2016) from Indiana University and MIMIC-CXR (Johnson et al. 2019) from the Beth Israel Deaconess Medical Center. The former dataset is relatively small with 7,470 chest X-Ray images and 3,955 radiology reports. The latter one

⁴It is worth noting that, the text generator learns to generate intermediate and final reports with different inputs. Compared with intermediate report generation, final report generation has an additional input (i.e., the intermediate report $\hat{\mathcal{Z}}$) and a different prompt.

is the largest public radiology dataset with 473,057 chest X-Ray images and 206,563 reports. We follow the experimental setup of previous studies (Li et al. 2018; Chen et al. 2020, 2021; Qin and Song 2022) by selecting findings sections and excluding samples without such sections for both datasets. We follow the dataset split in Li et al. (2018) for IU X-RAY and the official split of MIMIC-CXR. Table 1 reports the statistics of all datasets in terms of the numbers of radiographs, reports, patients, and average report length according to each split of the datasets.

3.2 Baselines and Evaluation Metrics

To verify our proposed approach, we try three baselines: the first uses the visual encoder and text generator to directly generate final reports, which is equivalent to the standard MiniGPT-4; the second and third baselines add I3 and C2FD on top of the first baseline, denoted as “+I3” and “+C2FD”, respectively. Note that we have two groups of such baselines according to whether MiniGPT-4 is fine-tuned on RRG data, with details of its fine-tune illustrated in the next subsection.

Following previous studies (Chen et al. 2020, 2021), we evaluate different approaches with two types of assessments, namely, natural language generation (NLG) metrics and clinical efficacy (CE) metrics. NLG metrics consist of BLEU (Papineni et al. 2002), METEOR (Denkowski and Lavie 2011) and ROUGE-L (Lin 2004). For CE metrics, CheXpert (Irvin et al. 2019) is utilized to label the generated reports and compare the results with ground truths in 14 different categories related to thoracic diseases and support devices, with precision, recall, and F1 used for evaluation.

DATA	MODEL	NLG METRICS						CE METRICS		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
IU X-RAY	ST (Vinyals et al. 2015)	0.216	0.124	0.087	0.066	-	0.306	-	-	-
	ATT2IN (Rennie et al. 2017)	0.224	0.129	0.089	0.068	-	0.308	-	-	-
	ADAATT (Lu et al. 2017)	0.220	0.127	0.089	0.068	-	0.308	-	-	-
	COATT (Jing, Xie, and Xing 2018)	0.455	0.288	0.205	0.154	-	0.369	-	-	-
	HRGR (Li et al. 2018)	0.438	0.298	0.208	0.151	-	0.322	-	-	-
	CMAS-RL (Jing, Wang, and Xing 2019)	0.464	0.301	0.210	0.154	-	0.362	-	-	-
	R2GEN (Chen et al. 2020)	0.470	0.304	0.219	0.165	-	0.371	-	-	-
	CA (Liu et al. 2021b)	0.492	0.314	0.222	0.169	0.193	0.381	-	-	-
	CMCL (Liu, Ge, and Wu 2021)	0.473	0.305	0.217	0.162	0.186	0.378	-	-	-
	PPKED (Liu et al. 2021a)	0.483	0.315	0.224	0.168	-	0.376	-	-	-
	R2GENCMN (Chen et al. 2021)	0.475	0.309	0.222	0.170	0.191	0.375	-	-	-
	R2GENRL (Qin and Song 2022)	<u>0.494</u>	<u>0.321</u>	<u>0.235</u>	<u>0.181</u>	<u>0.201</u>	<u>0.384</u>	-	-	-
	XRAYGPT (7B) (Thawkar et al. 2023)	0.177	0.104	0.047	0.007	0.105	0.203	-	-	-
	Ours (14.2B)	0.499*	0.323*	0.238*	0.184*	0.208*	0.390*	-	-	-
MIMIC -CXR	ST (Vinyals et al. 2015)	0.299	0.184	0.121	0.084	0.124	0.263	0.249	0.203	0.204
	ATT2IN (Rennie et al. 2017)	0.325	0.203	0.136	0.096	0.134	0.276	0.322	0.239	0.249
	ADAATT (Lu et al. 2017)	0.299	0.185	0.124	0.088	0.118	0.266	0.268	0.186	0.181
	TOPDOWN (Anderson et al. 2018)	0.317	0.195	0.130	0.092	0.128	0.267	0.320	0.231	0.238
	R2GEN (Chen et al. 2020)	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
	CA (Liu et al. 2021b)	0.350	0.219	0.152	0.109	0.151	0.283	-	-	-
	CMCL (Liu, Ge, and Wu 2021)	0.344	0.217	0.140	0.097	0.133	0.281	-	-	-
	PPKED (Liu et al. 2021a)	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
	R2GENCMN (Chen et al. 2021)	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
	R2GENRL (Qin and Song 2022)	0.381	0.232	0.155	0.109	0.151	<u>0.287</u>	0.342	0.294	0.292
	ITA (Wang et al. 2022a)	<u>0.395</u>	<u>0.253</u>	0.170	0.121	0.147	0.284	-	-	-
	WARMSTART (Aaron Nicolson 2022)	0.392	0.245	0.169	0.124	0.153	0.285	0.359	0.412	0.384
	RGRG (Tanida et al. 2023)	0.373	0.249	<u>0.175</u>	<u>0.126</u>	<u>0.168</u>	0.264	<u>0.461</u>	<u>0.475</u>	<u>0.447</u>
	XRAYGPT (7B) (Thawkar et al. 2023)	0.128	0.045	0.014	0.004	0.079	0.111	-	-	-
MED-PALM (562B) (Tu et al. 2023)	0.317	-	-	0.115	-	0.275	-	-	0.378	
Ours (14.2B)	0.402*	0.262*	0.180*	0.128*	0.175*	0.291*	0.465*	0.482*	0.473*	

Table 3: Comparisons of our approach with previous studies on the test sets of IU X-RAY and MIMIC-CXR with respect to NLG and CE metrics. The best and the second-best results are highlighted in boldface and underlines, respectively. For LLM-based methods (i.e., XRAYGPT, MED-PALM, and OURS), we illustrate the number of parameters in parentheses. “*” marks the results where the improvements are statistically significant over all baselines at $p \leq 0.05$ level.

3.3 Implementation Details

We use MiniGPT-4 with its default hyper-parameter settings (i.e., ViT-G version of vision transformer from EVA-CLIP (Fang et al. 2022) with 40 encoding layers, Q-Former (Li et al. 2023) with 12 layers, a linear projection layer corresponding to Eq. (5), and Vicuna (Chiang et al. 2023) (13B) as the text generator with its default 40 transformer layers). In order to obtain an enhanced baseline, we fine-tune MiniGPT-4 on the training set of public radiology benchmark MIMIC-CXR following the standard fine-tuning process, with the resulted model marked as MiniGPT-4 (FT) in following texts. For the public medical data used in I3, we randomly sampled 3,000 medical documents from PubMed dataset⁵. In addition, motivated by the practice of optimizing LLaMA-2 (Touvron et al. 2023b) by learning to rank instances in a binary form, we only retrieve a small number⁶ of related instances in I3, i.e., $M = 3$, which refers to that there are three instances retrieved from the training set and

⁵<https://pubmed.ncbi.nlm.nih.gov/pubmed>

⁶We try $M \in [1, 5]$ and adopt $M = 3$ with the best results.

three from the public medical corpora, respectively. For our full model, we train them on the training set of IU X-RAY and MIMIC-CXR with different hyper-parameter settings and use the one with the highest performance on the validation set. The batch sizes for IU X-RAY and MIMIC-CXR are set to 12. The weights to balance I3 and C2FD loss in Eq. (2) are set to $\beta_1 = 1$ and $\beta_2 = 1$, respectively. In training, we only update parameters in the linear projection layer in the visual encoder and Vicuna through AdamW (Kingma and Ba 2015) with learning rate set to 1×10^{-6} . Note that in inference, we only use visual encoding, intermediate report generation in I3 without RIR and other parts of CSR, and C2FD, to generate reports.

4 Results and Analysis

4.1 Overall Results

Experiment results of different models on the two benchmark datasets are reported in Table 2, with several observations. First, under different settings with original MiniGPT-4 and its fine-tuned version, our approach with I3 (i.e.,

	NLG METRICS					
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
I3 (TD)	0.339	0.182	0.120	0.084	0.126	0.238
I3 (PMD)	0.310	0.167	0.112	0.074	0.109	0.215

Table 4: I3 performance with only using training data (TD) or public medical data (PMD) as the retrieval source.

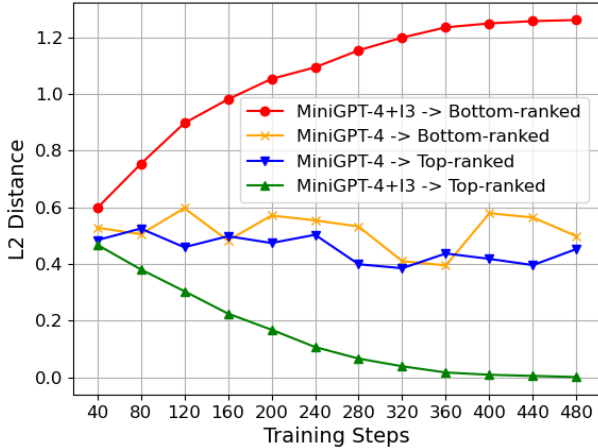


Figure 2: The curves of L2 distance between the representation of the reports from different models (i.e., MiniGPT-4 (FT)+I3 and MiniGPT-4 (FT)) and that of the top-ranked or bottom-ranked instances against training steps.

“+I3”) and C2FD (i.e., “+C2FD”) consistently outperforms all baselines on the test set of both datasets under all evaluation metrics, which demonstrates the effectiveness of our approach given the baselines, especially the MiniGPT-4 (FT), have already achieved outstanding performance. Second, the ablation of either I3 or C2FD leads to inferior performance compared with the full model, which presents that both components play essential roles in this task. Third, in most cases, model with I3 achieves better performance than the one with C2FD. The performance gap is much more significant when the models are equipped with the original MiniGPT-4 than the fine-tuned one, which indicates the power of I3 to bridge the domain gap and improve RRG, especially when the gap is rather large under the original MiniGPT-4 setting.

We further compare our approach (i.e., MiniGPT-4 (FT)+I3+C2FD) with existing state-of-the-art methods on the same datasets, where results are reported in Table 3 on both NLG metrics and CE metrics. Overall, our approach significantly outperforms all existing studies on both NLG and CE metrics, which further confirms the validity of our approach to enhance LLMs for RRG by learning from ranked instances and generating the final reports in a coarse-to-fine manner. Notably, our approach outperforms existing methods with medical domain LLMs (i.e., XRAYGPT (Thawkar et al. 2023) and MED-PALM⁷ (Tu et al. 2023))

⁷MED-PALM does not release the model weights and its RRG test set. Therefore, for fair comparisons, we approximate their settings to randomly curated 10 groups of test instances with the same size (i.e., 246 cases) as that used in MED-PALM. Our approach

	NLG METRICS					
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
INTER.	0.364	0.211	0.142	0.093	0.145	0.261
FINAL	0.402	0.262	0.180	0.128	0.175	0.291

Table 5: Comparison of intermediate (INTER.) and final reports that input to and output from C2FD w.r.t. NLG metrics.

with significant improvements.

4.2 Analysis

In this section, we analyze the effect of different components of our approach. Specifically for I3, we explore it separately with its two components, namely, RIR and CSR. Then, we investigate how C2FD performs and finally present the contribution of different components through a case study.

Effect of Related Instance Retrieval To investigate the effect of instance retrieval methods, we try alternative settings that only use the training data or public medical corpora as the retrieval sources in I3, as a comparison to that in our main experiment. Table 4 compares the results on MIMIC-CXR test set, where I3 (TD) and I3 (PMD) denote the aforementioned two settings, respectively. By comparing the results in Table 4 to “MiniGPT (FT)+I3” reported in Table 2, it is observed that I3 in our approach outperforms both I3 (TD) and I3 (PMD) with rather large margin, which confirms the validity of our design because single retrieval source contributes less differentiated instances for effectively ranking, therefore affects domain and task adaptation of the LLM to RRG. Moreover, the performance comparison between I3 (TD) and I3 (PMD) verifies that instances from the training data are more important than those ones irrelevant to the task, which further illustrates the feasibility and practicability of tuning LLM to RRG by our approach (especially using I3) with employing limited data.

Effect of Contrastive Semantic Ranking We analyze the changing of L2 distance between the representations of the intermediate reports and the ranked instances (i.e., $\|\mathbf{o} - \mathbf{o}'_m\|$) during training, so as to analyze how CSR performs. Figure 2 presents the curve of distance with respect to the training steps, where the red nodes stand for the distances between the intermediate report generated by MiniGPT-4 (FT)+I3 and the top-ranked instances, and green ones for that between the same report and bottom-ranked instances, respectively. As comparing references, blue/yellow nodes indicate the distances between the report generated by MiniGPT-4 (FT) and the top-/bottom-ranked instances, respectively. It is observed that, with more training steps, the distance (green dots) between the reports from MiniGPT-4+I3 and the top-ranked instances decreases and that (red dots) between the same report and the bottom-ranked instances increases, which indicates that our approach succeeds to enable the generator to produce texts that are closer to good references and separated from irrelevant information. On the contrary, MiniGPT-4 (FT) without the assistance achieves performance similar to the results reported in Table 3.


<p>Original Image</p>  <p>Gold Standard: Lung volumes are low. Mediastinal and hilar contours are unremarkable. The heart is mildly enlarged. Streaky opacities in the lung bases likely reflect areas of atelectasis. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.</p>	<p>MiniGPT-4: Key Findings: * A large mass in the right lung field, likely a malignancy * Irregularly shaped mass with an irregular border...</p>
	<p>MiniGPT-4 (FT): The given radiology image is a frontal chest radiograph of an adult male. The image is oblique, with no abnormalities visible. The lungs are well aerated and the heart is in its normal position. The mediastinum is not enlarged and there is no evidence of any infiltrates or consolidation.</p>
	<p>MiniGPT-4 (FT)+I3: Lung volumes are normal. The heart size is within normal limits. The thoracic aorta is widened but elongated. No local abnormality is identified. The cardiac and mediastinal silhouettes are unremarkable. No abnormalities in local contours are identified. No pleural effusion and pneumothorax is seen.</p>
	<p>MiniGPT-4 (FT)+C2FD: Frontal chest radiograph is presented. The image is normal and unremarkable, with no evidence of vascular abnormalities. The chest radiograph shows moderate opacities in both lungs, consistent with chronic bronchitis. The heart is mildly enlarged. Mediastinum contours are clear and there is no pleural effusion or focal consolidation.</p>
	<p>MiniGPT-4 (FT)+I3+C2FD: There are low lung volumes. Heart size is mildly enlarged. Mediastinal contours are unremarkable. Patchy atelectasis are seen at the lung bases. There is mild pulmonary vascular congestion without overt interstitial edema. No pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.</p>

Figure 3: An illustration of the reports generated by different models with a given input radiograph, where the gold standard is also presented for reference. Medical terms shared by the gold standard and model outputs are highlighted in the same color. Correct and incorrect content in the generated reports are highlighted in green and red background, respectively.

tance of I3 is unable to distinguish related reports from the unrelated ones, which is not capable of aligning LLMs to the task and thus obtains inferior performance.

Effect of Coarse-to-Fine Decoding To investigate how C2FD performs, we run our full model and extract the reports input (intermediate) to and output (final) from C2FD. Table 5 compares NLG results of those reports on MIMIC-CXR test set, where significant improvements are observed on final reports, confirming the effectiveness of C2FD in refining the intermediate reports with precise and coherent elevation when an appropriate model design is applied.

Case Study To further qualitatively investigate how our approach bootstraps LLMs for RRG, we perform a case study on the output reports of different models with the same input chest X-ray image chosen from MIMIC-CXR. Figure 3 shows the results, with several observations from different perspectives drawn as follows. The original MiniGPT-4 fails to handle RRG and generates reports that contain few medical terms. On the contrary, MiniGPT-4 (FT) generates a much better report with more relevant medical terms (e.g., “heart”, “mediastinum”) compared to the original MiniGPT-4. However, the generated reports still include irrelevant descriptions (e.g., “This image is oblique.”) owing to less alignment to the task, which is also found in MiniGPT-4 (FT)+I3 although it produces reports that are well patternized and better aligned to the medical domain. MiniGPT-4 (FT)+C2FD produces more informative and precise reports than that without C2FD, where some missing diagnoses from MiniGPT-4 (FT) are resolved by the C2FD process (e.g., “There is no pleural effusion”). Finally, MiniGPT-4 (FT)+I3+C2FD generates the best report over all other models, whose problems are all alleviated to some extent, thus confirms the superiority of the proposed approach.

5 Related Work

Clinical medicine has raised increasing attention nowadays (Wu et al. 2019; Tian et al. 2020; Song et al. 2020). Particularly, RRG is a challenging application, requiring to generate long text in the medical domain. In doing so, some studies try to leverage useful visual and textual features to improve RRG, e.g., regional visual features (Tanida et al. 2023), report templates (Li et al. 2018), and structure-level descriptions (Wang et al. 2022a). From another aspect, some studies focus on improving cross-modal alignment through co-attention (Jing, Xie, and Xing 2018), memory networks (Chen et al. 2020, 2021), and reinforcement learning (Qin and Song 2022), to better matching different information to guide the generation process. Recently, with the extraordinary generation ability of LLMs (Touvron et al. 2023a; Gan et al. 2023; Yuanhe Tian 2023), recent studies have applied LLMs to multimodal scenarios (Li et al. 2023; Zhu et al. 2023; Liu et al. 2023), including applications in the medical domain (Thawkar et al. 2023; Tu et al. 2023). Compared with them, our approach offers an alternative solution to improve RRG, where we start from general domain LLM (i.e., MiniGPT-4) and design special learning processes to bootstrap it for RRG with limited in-domain data for domain adaptation and generation optimization.

6 Conclusion

In this paper, we propose to bootstrap LLMs for RRG, where I3 and C2FD are proposed to align LLMs with the medical domain and improve report generation, respectively. Experimental results demonstrate our superiority to current state-of-the-art models on IU X-Ray and MIMIC-CXR with analysis further conducted to verify its validity. So that the fine-tuned LLM and our designed components in this work offer a new reference framework for future RRG studies. Notably, our approach provides a practical paradigm of adapting general domain LLMs to applications in specific domains, which reveals the potential of extending the approach to other tasks.

References

- Aaron Nicolson, B. K., Jason Dowling. 2022. Improving Chest X-Ray Report Generation by Leveraging Warm-Starting. arXiv:2201.09405.
- Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 6077–6086.
- Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; Das-Sarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.
- Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5904–5914. Online.
- Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1439–1449. Online.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.
- Demner-Fushman, D.; Kohli, M.; Rosenman, M.; Shooshan, S.; Rodriguez, L.; Antani, S.; Thoma, G.; and McDonald, C. 2016. Preparing A Collection of Radiology Examinations for Distribution and Retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.
- Denkowski, M.; and Lavie, A. 2011. METEOR 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 85–91. Edinburgh, Scotland.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Hounsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*, 1–21.
- Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2022. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. *arXiv preprint arXiv:2211.07636*.
- Gan, R.; Wu, Z.; Sun, R.; Lu, J.; Wu, X.; Zhang, D.; Pan, K.; Yang, P.; Yang, Q.; Zhang, J.; and Song, Y. 2023. Ziya2: Data-centric Learning is All LLMs Need. *arXiv preprint arXiv:2311.03301*.
- Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilicus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpan-skaya, K.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C. P.; Patel, B. N.; Lungren, M. P.; and Ng, A. Y. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv:1901.07031.
- Jing, B.; Wang, Z.; and Xing, E. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6570–6580. Florence, Italy.
- Jing, B.; Xie, P.; and Xing, E. 2018. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2577–2586. Melbourne, Australia.
- Johnson, A. E. W.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; ying Deng, C.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, A Large Publicly Available Database of Labeled Chest Radiographs. arXiv:1901.07042.
- Kingma, D.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 1–15. San Diego, CA, USA.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.
- Li, Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2018. Hybrid Retrieval-generation Reinforced Agent for Medical Image Report Generation. In *NeurIPS*, 1537–1547.
- Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain.
- Liu, C.; Tian, Y.; and Song, Y. 2023. A Systematic Review of Deep Learning-based Research on Radiology Report Generation. arXiv:2311.14199.
- Liu, F.; Ge, S.; and Wu, X. 2021. Competence-based Multimodal Curriculum Learning for Medical Report Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3001–3012. Online.
- Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021a. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. arXiv:2106.06963.
- Liu, F.; Yin, C.; Wu, X.; Ge, S.; Zhang, P.; and Sun, X. 2021b. Contrastive Attention for Automatic Chest X-ray Report Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 269–280. Online.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. arXiv:2304.08485.
- Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. arXiv:1612.01887.

- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training Language Models to Follow Instructions with Human Feedback. arXiv:2203.02155.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA.
- Qin, H.; and Song, Y. 2022. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, 448–458. Dublin, Ireland.
- Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical Sequence Training for Image Captioning. arXiv:1612.00563.
- Song, Y.; Tian, Y.; Wang, N.; and Xia, F. 2020. Summarizing Medical Conversations via Identifying Important Utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, 717–729.
- Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *CVPR*, 7433–7442.
- Thawkar, O.; Shaker, A.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. S. 2023. XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models. arXiv:2306.07971.
- Tian, Y.; Shen, W.; Song, Y.; Xia, F.; He, M.; and Li, K. 2020. Improving Biomedical Named Entity Recognition with Syntactic Information. *BMC Bioinformatics*, 21: 1471–2105.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.
- 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.; and Scialom, T. 2023b. Llama 2: Open Foundation and Fine-tuned Chat Models. arXiv:2307.09288.
- Tu, T.; Azizi, S.; Driess, D.; Schaeckermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; Mustafa, B.; Chowdhery, A.; Liu, Y.; Kornblith, S.; Fleet, D.; Mansfield, P.; Prakash, S.; Wong, R.; Virmani, S.; Sertur, C.; Mahdavi, S. S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Singhal, K.; Florence, P.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Generalist Biomedical AI. arXiv:2307.14334.
- Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and Tell: A Neural Image Caption Generator. arXiv:1411.4555.
- Wang, L.; Ning, M.; Lu, D.; Wei, D.; Zheng, Y.; and Chen, J. 2022a. An Inclusive Task-aware Framework for Radiology Report Generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, 568–577. Cham. ISBN 978-3-031-16452-1.
- Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022b. Med-CLIP: Contrastive Learning from Unpaired Medical Images and Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3876–3887. Abu Dhabi, United Arab Emirates.
- Wu, Z.; Song, Y.; Huang, S.; Tian, Y.; and Xia, F. 2019. WTMed at MEDIQA 2019: A Hybrid Approach to Biomedical Natural Language Inference. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 415–426. Florence, Italy.
- Yuanhe Tian, Y. S. J. Z. Y. Z., Ruyi Gan. 2023. ChiMed-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences. *arXiv preprint arXiv:2311.06025*.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models. arXiv:2304.10592.