Overview of the MEDIQA-M3G 2024 Shared Task on Multilingual Multimodal Medical Answer Generation

Wen-wai Yim, Asma Ben Abacha
Microsoft Health AI
{yimwenwai,abenabacha}@microsoft.com

Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen
University of Washington
{velvinfu,zhaoyis,fxia,melihay}@uw.edu

Martin Krallinger
Barcelona Supercomputing Center
martin.krallinger@bsc.es

Abstract

Remote patient care provides opportunities for expanding medical access, saving healthcare costs, and offering on-demand convenient services. In the MEDIQA-M3G 2024 Shared Task, researchers explored solutions for the specific task of dermatological consumer health visual question answering, where user generated queries and images are used as input and a free-text answer response is generated as output. In this novel challenge, eight teams with a total of 48 submissions were evaluated across three language test sets. In this work, we provide a summary of the dataset, as well as results and approaches. We hope that the insights learned here will inspire future research directions that can lead to technology that deburdens clinical workload and improves care.

1 Introduction

Driven by long patient wait times, high medical costs and physician burnout, remote patient care delivery (e.g. e-visits, e-mails) provides a cost effective solution that lowers facility expenditures, allows flexible schedules for both clinicians and patients, and expands health care access(Bishop et al., 2024). The trend, already in motion due to the maturation of telecommunication technologies and the proliferation of health portals, was massively accelerated by the onset of the global COVID 19 epidemic in 2019. Five years later, today, while remote technologies allow for the conveniences of patient care delivered conveniently from one's own home, this poses new challenges for providers who need to meet the new demand, where patients can request services at any time of the day, creating a perception of "never-ending work"(Sinsky et al., 2024).

Automatic response generation may alleviate doctor burden by providing suggestions when answering patient queries, speeding up response throughput. In this work, we present the MEDIQA 2024 Multilingual & Multimodal Medical Answer Generation (M3G) Shared Task, which is focused on the problem of multimodal answer generation in the space of dermatology, evaluated in multiple languages (English, Chinese, and Spanish). Specifically, a health related question along with one or more images is posed; the expected task is to generate an appropriate answer response.

Previous editions of the MEDIQA shared tasks have featured radiology-related visual question answering(Lau et al., 2018; Ben Abacha et al., 2019) and text-only consumer health answer generation(Ben Abacha et al., 2017). Other prior work in medical VQA includes images in the space of pathology and GI-tract (He et al., 2021; Hicks et al., 2023). Meanwhile, previous work in dermatological image classification focused on image-only input and multi-class classification(Daneshjou et al., 2021; Groh et al., 2021). This is the first shared task to incorporate visual question answering for user generated health queries and images.

2 Task

2.1 Description

In this task, participants were given textual inputs which may include clinical history and a query, along with one or more associated images. The task objective consisted of generating a relevant textual answer response. An example instance is shown in Table 1.

The training set contained multiple possible gold standard responses. Each response included information related to its author validation level (e.g. real-id verified, medical doctor verified) and a ranking based on their platform contribution from 0-8 levels, the higher the better. English and Spanish translations were automatically generated from original Chinese (in simplified Chinese characters) using GPT4.

In the validation and test sets, each text response

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 581–589

Table 1: Example from the DermaVQA IIYI data subset. The original posts are in Chinese, which are translated into English and Spanish by GPT4 (if they are in the training set) or medical translators (otherwise).

Query	Responses
 帮忙诊断一下:三个月前出现如下图，自己用达克能宁喷雾两个月无明显效果，之后去乡村诊所，医生指导用鸡眼膏，之后出现变红变多，请帮忙诊断下 Please help with the diagnosis: Three months ago, the condition shown in the picture below appeared. The patient used Daknening spray for two months without any noticeable effect. Afterwards, they went to a rural clinic, where the doctor advised them to use corn ointment. Subsequently, the condition turned red and worsened. Please help with the diagnosis. Por favor, ayude con el diagnóstico : Hace tres meses, apareció la condición mostrada en la imagen de abajo. El paciente utilizó el spray Daknening durante dos meses sin ningún efecto notable. Posteriormente, acudió a una clínica rural, donde el médico le aconsejó que utilizara pomada de maíz. Posteriormente, la condición se volvió roja y empeoró. Por favor, ayude con el diagnóstico.	RESPONSE1: 是鸡眼。 It's a corn. Es un callo. RESPONSE2: 考虑：跖疣 Consideration: Plantar wart Consideración: ¿Verruga plantar RESPONSE3: 是跖疣，不是鸡眼，激光治疗。 It's a plantar wart, not a corn. Laser treatment is recommended. Es una verruga plantar, no un callo. Se recomienda el tratamiento con láser.

was also given a human rating for completeness and whether an answer is one that was the most frequent. The rating guide is as follows: 0.0 for no, 0.5 for partial and, 1.0 for yes. English and Spanish versions were human translated by medical translators.

2.2 Dataset

The dataset here was constructed by using content from a Chinese online medical platform 爱爱医[1] for posts related to dermatological problems. In the platform, users may post a question with images; doctors on the platform may respond. Thus, in our dataset, in each instance, the input is the original query and images provided by the original poster. The answer is the set of answers provided by medical experts who responded to the query.

Encounters were filtered out if it met at least one of the following exclusion criteria: (a) images that included identifying features (e.g. full faces), (b) no medical answers were given, (c) queries were not seeking information (e.g. "look at my tatoo"), and (d) images contained annotations (e.g. drawn arrows). Train/validation/test sets included 842/56/100 instances, respectively. Table 2 shows summary statistics of the data. A query can involve multiple anatomic locations and medical topics (calculated by counting terms identified using QUICKUMLS(Soldaini and Goharian, 2016) on the English). The test set required at least two responses.

The data here used a subset of the DermaVQA dataset, for which the full description can be found in (Yim et al., 2024b).

2.3 Evaluation

We evaluate the system responses by comparing with the multiple gold standard responses per query. We used relevant multi-reference metrics/variants including:

deltaBLEU. A variant of SacreBLEU developed for response generation, a case in which many diverse gold standard responses are possible (Galley et al., 2015). The metric incorporates human-annotated quality rating and assigns higher weights to n-grams from responses rated to be of higher quality. The authors have shown this method produces higher correlation with human rankings compared to previous BLEU metrics. In our system, we assign response weights according to four criteria: (a) if user expertise level is 4 or above (out of 9), (b) if user is formally validated as a medical doctor by the platform, (c) if the response answer is the most

[1] iiyi.com

Table 2: DermaVQA IIYI Subset Data Characteristics. (encs=encounters, encs-x img=number of encounters with x images, encs-x resp=number of encounters with x responses)

	TRAIN	VALID	TEST	TOTAL
N	842	56	100	998
IMAGES				
total count	2473	157	314	2944
mean count	2.9	2.8	3.1	2.95
encs-1 img	196	11	18	225
encs-2 img	233	22	30	285
encs-3 img	171	11	18	200
encs->=4 img	242	12	34	288
RESPONSES				
total count	5871	417	926	7214
mean count	7.0	7.4	9.3	7.2
encs-1 resp	66	0	0	66
encs-2 resp	80	6	5	91
encs-3 resp	100	4	6	110
encs->=4 resp	596	46	89	731
LENGTH (words/char)				
per query(en/es/zh)	80.4/81.8/89.0	75.0/71.9/79.0	76.0/74.3/81.0	79.6/80.5/87.6
per response(en/es/zh)	11.9/12.7/16.4	14.9/15.2/19.6	10.8/10.7/14.0	11.9/12.6/16.3
MEDICAL TOPICS				
Diagnosis	610	196	137	695
Tests	39	10	13	46
Treatments	494	123	104	567
LOCATIONS				
Arm region	162	6	19	187
Back region	85	10	9	104
Chest/Abdomen region	107	4	13	124
Foot region	129	8	15	152
Hand region	221	19	31	271
Head region	178	12	13	203
Leg region	198	12	21	231
UNSPECIFIED	161	9	25	195

frequent answer, and (d) if the response answers the query completely. The former two were manually assigned to the validation and test sets by two NLP scientists. The test set was double-reviewed. Out of a 0.0-1.0 scale, if (d) is not met, the score is discounted to 0.9; for the other 3 criteria, 0.1 is discounted for every missing element to reach the final weight.

BERTScore. BERTScore[2](Zhang et al., 2019) averages the maximum word embedding similarity scores between two texts based on BERT embeddings. This metric has been shown to work well on a variety of tasks, including image captioning and machine translation. The maximum was taken over multiple candidate scores.

MEDCON. For the English and Chinese datasets, we also used a medical information extraction based metric, here known as MEDCON. The metric uses QuickUMLS[3] to identify medical concepts in conjunction with an in-house llama-based assertion classifier. Concepts identified by QuickUMLS are normalized according to a curated concept map. Precision, recall, and F1 were calculated based on combined concept and assertion statuses. The maximum was taken over multiple candidate scores.

Evaluation code can be found in our github repo[4].

[2]github.com/Tiiiger/bert_score

[3]github.com/Georgetown-IR-Lab/QuickUMLS

[4]https://github.com/wyim/MEDIQA-M3G-2024

Table 3: Participant Information

TEAM	AFFLIATION	LANGUAGES
Intel-PA	United Kingdom	en,zh
LTRC-IIITH (Thomas et al., 2024)	International Institute of Information Technology - Hyderabad, India	en
MediFact (Saeed, 2024)	Pakistan	en,es,zh
MIPAL (Kim et al., 2024)	Seoul National University / Republic of Korea	en
NEUI (García and Lithgow-Serrano, 2024)	The Dalle Molle Institute for Artificial Intelligence (IDSIA), Switzerland	en
IKIM (Bauer et al., 2024)	Institute for AI in Medicine (IKIM), University Hospital Essen, Germany	en,es,zh
UMass-BioNLP (Vashisht et al., 2024)	University of Massachusetts Amherst, USA	en
WangLab (Toma et al., 2024)	University of Toronto, Canada	en

3 Baselines

For reference, we provide two baselines based on vision-language models and an additional translation step: ***Baseline1:*** GPT4V and GPT4(gpt, 2024), ***Baseline2:*** Gemini-Pro-Vision and Gemini-Pro(gem, 2024) are used.

The prompt for generating content was " 请以专业人士皮肤科医生来回答，答案只限17字，不用加客气的说话。 {post_title}: {post_content}" (Translation: Please answer as a professional dermatologist, answer limited to 17 characters, do not include pleasantries). The English and Spanish baselines were translations of the Chinese output. The prompt for the machine translation was "Instructions: Translate the following medical text faithfully from Chinese into {TARGET_LANGUAGE}."

4 Official Results

4.1 Participating teams

The shared task included 52 of registered participants. The final number of teams that submitted runs was 8 teams, with a total of 48 submissions. Participating teams came from various regions including Europe (3), North America (2), South Asia (2), and East Asia (1). The number of teams and submissions were 8 and 36 for English, 3 and 12 for Chinese, and 3 and 6 for Spanish. We limited the number of runs to 10. Details of the participating teams are shown in Table 3. deltaBLEU was used for official ranking.

4.2 Approaches and Results

Tables 4, 5, 6 detail the results for the English, Chinese, and Spanish test sets respectively. The BLEU scores ranged between 0.231-12.855, 2.171-7.053, and 0.446-1.355 for English, Chinese, and Spanish test sets. It is notable that the magnitude of scores for both Chinese and Spanish test sets did not vary widely, the top three scores for English was significantly higher than other systems with the difference between the third best system and fourth at 7 BLEU points. BERTScore had higher ranges for English (0.800-0.886), and lower ranges for Chinese (0.685-0.764) and Spanish (0.764-0.818). In general the MEDCON scores were low, with the highest number at 0.287.

Fine-tuned Vision-language Models Systems: Three teams–Team MIPAL, IKIM, and LTRC-IIITH–relied fine-tuning visual-language models. The models included MedVInT(Zhang et al., 2023) and LLaVA(Liu et al., 2023), LLaVA-Med(Li et al., 2023), ViLT(Kim et al., 2021), respectively. The score variation, ranging from 0.457 to 3.827 BLEU suggests the combination of model, prompts, and fine-tuning strategy lead to large differences in results.

Pre-trained Vision-language Systems: As multiple submissions were allowed, the previous teams also submitted non-fined-tuned model outputs as shown in the FINE_TUNED columns of Tables 4, 5, 6. For non open models, in one submission, Team WangLab experimented with Claude3 Opus(ant, 2024), using two calls - one for candiate generation another for a final response, with competitive results. Likewise, the UMass-BioNLP used pre-trained models without fine-tuning in a multi-step fashion. The team first employed GPT-4/GPT-4-Vision(Wu et al., 2023) to generate initial hypotheses; secondly they generated image descriptors from the disease candidates of the previous step. Afterwards, they selected possible diagnosis by comparing image descriptors similarities of the disease candidates and that of the image descriptors from

Table 4: Results (English) - Top 3 Results per Team

RANK	TEAM	FINE_TUNED	MODELS	deltaBLEU	BERTScore	MEDCON
1	WangLab	FALSE	clip	12.855	0.882	0.222
2	WangLab	FALSE	Claude. based prompt engineering	12.159	0.886	0.287
3	WangLab	FALSE	fine-tuned clip	11.979	0.862	0.125
4	MIPAL	TRUE	PMC-VQA(PMC-CLIP, PMC-LLaMA)	3.827	0.872	0.139
5	MIPAL	TRUE	PMC-VQA(PMC-CLIP, PMC-LLaMA)	3.263	0.872	0.139
6	MIPAL	TRUE	PMC-VQA(PMC-CLIP, PMC-LLaMA)	3.263	0.872	0.139
7	IKIM	TRUE	llava-med + mixtral-instruct	2.662	0.858	0.123
8	IKIM	TRUE	llava-med, mixtral	2.662	0.858	0.123
9	NEUI	FALSE	Phi1	2.133	0.850	0.131
10	Intel-PA	FALSE	BLIP2	1.758	0.852	0.155
11	Intel-PA	FALSE	Intel-PA-run8	1.505	0.849	0.180
12	UMass-BioNLP	FALSE	GPT4	0.923	0.852	0.159
13	UMass-BioNLP	FALSE	GPT4	0.823	0.851	0.131
14	MediFact	TRUE	VGG16-CNN-SVM	0.717	0.842	0.148
15	Intel-PA	FALSE	BLIP2	0.711	0.837	0.086
16	UMass-BioNLP	FALSE	GPT4	0.670	0.821	0.158
17	NEUI	FALSE	Phi1	0.595	0.851	0.205
18	MediFact	TRUE	VGG16-CNN-SVM	0.588	0.845	0.163
19	MediFact	FALSE	BART, SVM, TF-IDF	0.588	0.838	0.054
20	IKIM	FALSE	llava med on chinese data + translation	0.554	0.860	0.057
21	LTRC-IIITH	FALSE	Vision-and-Language Transformer (ViLT) model - dandelin/vilt-b32-mlm	0.457	0.829	0.016
22	neui	TRUE	Phi1	0.231	0.810	0.065
-	baseline1	FALSE	GPT4	0.813	0.867	0.083
-	baseline2	FALSE	GEMINI	1.094	0.800	0.157

Table 5: Results (Chinese) - All Results

RANK	TEAM	FINE_TUNED	MODELS	deltaBLEU	BERTScore	MEDCON
1	IKIM	TRUE	llava-med, mixtral-instruct	7.053	0.764	0.067
2	IKIM	FALSE	llava-med, mixtral	7.053	0.764	0.074
3	IKIM	FALSE	llava-med, Biomistral	7.053	0.764	0.060
4	Intel-PA	FALSE	–	6.976	0.756	0.031
5	Intel-PA	FALSE	BLIP2	6.976	0.756	0.029
6	Intel-PA	FALSE	–	5.166	0.757	0.017
7	Intel-PA	FALSE	BLIP2	5.032	0.741	0.027
8	MediFact	TRUE	VGG16-CNN-SVM	4.503	0.763	0.106
9	MediFact	TRUE	VGG16-CNN-SVM	4.503	0.763	0.105
10	Intel-PA	FALSE	BLIP2	4.073	0.731	0.036
11	Intel-PA	FALSE	BLIP2	2.426	0.712	0.015
12	MediFact	FALSE	BART, SVM, TF-IDF	2.171	0.707	0.075
-	baseline1	FALSE	GPT4	7.025	0.735	0.016
-	baseline2	FALSE	GEMINI	9.311	0.685	0.107

Table 6: Results (Spanish) - All Results

RANK	TEAM	FINE_TUNED	MODELS	deltaBLEU	BERTScore
1	IKIM	TRUE	llava-med, mixtral	1.355	0.818
2	NEUI	FALSE	Phi1	0.974	0.814
3	NEUI	FALSE	Phi1	0.974	0.814
4	MediFact	TRUE	VGG16-CNN-SVM	0.918	0.806
5	MediFact	TRUE	VGG16-CNN-SVM	0.823	0.809
6	MediFact	FALSE	BART, SVM, TF-IDF	0.446	0.802
-	baseline1	FALSE	GPT4	0.979	0.822
-	baseline2	FALSE	GEMINI	1.355	0.764

the encounter images outputted by GPT-4-Vision.

Multi-step Mixed Model Systems: The teams, Teams Intel-PA, NEUI, MediFact, and WangLab, experimented with a series of multiples steps using both fine-tuned and pre-trained models in a pipeline. Team Intel-PA uses a BLIP2 model, taking the output layer and combining word embeddings. These combined vectors were then fed to a large language model for text generation. Team NEUI used a fine-tuned visual language model, Moondream (https://moondream.ai/), to generate candidates. Then candidates were given as input into a BioMistral-7B-DARE(Labrak et al., 2024) to produce the final output. Team MediFact experimented with various image embedding methods, e.g. CLIP and VGG16, with a prediction task to classify a training answer response label using an SVM. The previous output combined with the query information was then fed into a reading comprehension model, Medical-QA-deberta-MRQA-COVID-QA(mrq, 2024), to generate an intermediate output. The final response is chosen by leveraging CLIP and finding the highest similarity of the image and QA output to a trained response. Google translator was used to generate the Chinese and Spanish versions. Team WangLab experimented embedding images using a fine-tuned CLIP model. The highest similarity to the test set was retrieved; the label selected from multiple gold responses in the test set was determined using GPT4. Finally, the retrieved labels were post-processed to an expected sentence format.

Multilingual Generation Approaches

Three patterns emerged for handling of multiple languages: (a) separate fine-tuning for each language, (b) prompt-adjustment as in Team NEUI, e.g. instructing output to be in Spanish, (c) a separate machine translation step as in Team IKIM, MediFact.

While Team IKIM fine-tuned on the Chinese dataset, then translated to English and Spanish separately using a Mixtral-8x7B-instruct model(Jiang et al., 2024); Team NEUI focused on English, translating to Spanish. The performance gap between IKIM and NEUI in English was at 0.529 BLEU, and 0.37 BLEU in Spanish. Though they used different systems, the relative scoring gap was preserved, suggesting that the two methods (b) and (c) are comparable.

The comparative effect of fine-tuning on automatically translated text prior to training versus using the original language and translating after generation requires further study.

4.3 Discussion and Related Work

The baseline systems using out-of-the box GPT-4-Vision and Gemini-Pro-Vision showed highly competitive performance for its original Chinese language at 7.025 and 9.311 BLEU (Table 5). However, this performance drops considerably when the same text is translated to English and Spanish; then evaluated on those test sets. Part of this drop may be due to automatic translation error, however this difference can also be partly attributed to the n-gram treatment of Chinese characters compared to latin words; which allows more partial credit. BERTScores were more stable across other languages, however are the comparatively higher compared to other metrics. MEDCON, a relatively simple, but strict metrics showed lower scores across datasets, suggesting much room for further improvement.

Although scores here are modest compared to previous Visual Question Answering (VQA) tasks. On further examination, this difference is due to the

nature of expected answers. Prior VQA datasets have question types with 1 or 2 fixed expected categorical responses. In fact, except for one work, all previous VQA tasks report accuracy as a metric. For the three cases of prior work that also report BLEU, average answer length was around 2 words. BLEU-3 scores for PathVQA were at most 17.4, even with at least half the corpus including a yes/no question type. BLEU ranges for the VQA-RAD, with more open-ended questions, achieved scores ranging from a modest 0.0058 to 0.1047 BLEU. This is consistent with recent studies which have shown that when queries are converted from a closed question-answering setting, e.g. multiple choice, to an open question-answering setting, this leads to significant degradations in performance, as much as 20%(Yim et al., 2024a).

A comparison with prior dermatological image classification tasks with user generated images also lend a helpful landscape. In Glock et al(Glock et al., 2021), with two classification categories an accuracy 95% was achieved; however for a dataset like SD-128, 128 categories, accuracy was at 52%. In a direct comparison, the authors of the Fitzpatrik 17k dataset study found a 20% accuracy when using 114 skin conditions which rises to 62% when simplifying to three categories (non-neoplastic, benign, and malignant)(Groh et al., 2021). As our gold standard responses were not generated using a fixed vocabulary, all the possible types and subtypes of diagnosis, treatments, and recommendations contributed to the difficulty of the task.

5 Conclusion and Future Work

Open-ended consumer health visual question answering remains a challenging problem. This shared tasks highlights several areas for future work.

One aspect is related to the generation of a dermatology common problem gold standard. Here we used a dataset with multiple references, some with varying opinions. For the dermatological specialty, a true gold standard with pathological lab confirmation is difficult to obtain in real life. This reflects the realities of current healthcare technology and costs – biological sampling and assays are only reserved for the most severe cases. Thus, datasets with biopsy observations are highly biased towards problems suspected to be malignant; whereas the plethora of other common-place maladies will remain unconfirmed. Textbook images and diagnosis labels, on the other hand, will not include user-generated queries. This is a non-trivial hurdle if an unequivocal dermatological VQA gold standard beyond medical doctor opinion is to be achieved. Furthermore, the dataset here limits responses to queries to a single turn - however multiple turns are necessary for clarification purposes in real clinical settings.

Another future direction is the development of mature evaluation methods when multiple references of varying quality is available. In past TREC competitions, one evaluation strategy included the employment of expert humans who would annotate each participant system based on answer quality(Ben Abacha et al., 2017). Ratings include categories: (a) Correct and Complete Answer, (b) Correct but Incomplete, (c) Incorrect but Related, and (d) Incorrect. In this task, we sought to incorporate this automatically in terms of weighing response answers for BLEU. However, although this sidesteps a need for a human expert to rate each system output, this method still relies on some human annotation of the gold standard instances. As well, the final scoring depends heavily on the quality and variety of existing answers; this leaves room for metric exploitation given the data biases. For example, on observation of the test set, although responses may include a variety of responses including recommended diagnosis, treatments, and test suggestions; since most responses at least give a diagnosis, it is advantageous to optimize for a short disease response instead of try to add more details and possibly incur penalties with an incorrect suggestion. Furthermore, mentioned medical concepts may have hierarchical relations with those the gold standard for current metrics do not take into account for well. For example, atopic dermatitis is equivalent to eczema and is a subtype of dermititis – however, eczema is not the same as contact dermatitis. Depending on the available combinations of gold responses, the same system output may receive different scores.

In this shared task, a variety of solutions were explored to provide solutions for the dermatological VQA. We hope that the benchmarks provided here, the insights from different systems, and the identified methological problems will inspire future research directions.

Limitations

The paper does not cover all types of possible methods and models for the generation of dermatological consumer health queries. The challenge datasets are limited in terms of size and coverage of diseases, treatments, and question types. The scope of the dataset only covers single turn responses. Further experiments and evaluations are needed to validate the best performing methods on other datasets and scenarios.

Acknowledgments

We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP organizers for their feedback and support for the MEDIQA-M3G 2024 shared tasks. We also thank Eulàlia Farré-Maduell and our diverse annotation team for preparing the data in time for the challenge and all the participating teams who contributed to the success of these shared tasks through their interesting approaches and experiments and strong engagement.

The authors acknowledge the support from the Spanish Ministerio de Ciencia e Innovación (MICINN) under project PID2020-119266RA-I00 and BARITONE (TED2021-129974B-C22). This work is also supported by the European Union's Horizon Europe Co-ordination & Support Action under Grant Agreement No 101080430 (AI4HF) as well as Grant Agreement No 101057849 (DataTool4Heartproject).

References

2024. Anthropic. https://www.anthropic.com/news/claude-3-family. Accessed: 2024-04-24.

2024. Gemini models. https://ai.google.dev/gemini-api/docs/models/gemini. Accessed: 2024-04-24.

2024. Gpt-4 turbo and gpt-4. https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4. Accessed: 2024-04-24.

2024. Medical-qa-deberta-mrqa-covid-qa. https://huggingface.co/longluu/Medical-QA-deberta-MRQA-COVID-QA. Accessed: 2024-04-24.

Marie Bauer, Amin Dada, Constantin Marc Seibold, and Jens Kleesiek. 2024. Ikim at mediqa-m3g 2024: Multilingual visual question-answering for dermatology through vlm fine-tuning and llm translations. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC 2017*.

Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.

Tara F. Bishop, Matthew J. Press, Jayme L. Mendelsohn, and Lawrence P. Casalino. 2024. Electronic communication improves access, but barriers to its widespread adoption remain. *Health affairs (Project Hope)*, 32(8):10.1377/hlthaff.2012.1151.

Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A. Novoa, Melissa Jenkins, Veronica Rotemberg, Justin M. Ko, Susan M. Swetter, Elizabeth E. Bailey, Olivier Gevaert, Pritam Mukherjee, Michelle Phung, Kiana Yekrang, Bradley Fong, Rachna Sahasrabudhe, James Zou, and Albert S. Chiou. 2021. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 445–450, Beijing, China. Association for Computational Linguistics.

Ricardo Omar Chàvez García and Oscar William Lithgow-Serrano. 2024. Neui at mediqa-m3g 2024: Medical vqa through consensus. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Kimberly Glock, Charlie Napier, Todd Gary, Vibhuti Gupta, Joseph Gigante, William Schaffner, and Qingguo Wang. 2021. Measles rash identification using transfer learning and deep convolutional neural networks. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3905–3910.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2021. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 708–718, Online. Association for Computational Linguistics.

Steven Hicks, Andrea M. Storås, Pål Halvorsen, Thomas de Lange, M. Riegler, and Vajira Lasantha Thambawita. 2023. Overview of imageclefmedical 2023 - medical visual question answering for gastrointestinal tract. In *Conference and Labs of the Evaluation Forum*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Hyeonjin Kim, MIN KYU KIM, Jae Won Jang, KiYoon Yoo, and Nojun Kwak. 2024. Team mipal at mediqa-m3g 2024: Large vqa models for dermatological diagnosis. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *Preprint*, arXiv:2102.03334.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Nadia Saeed. 2024. Medifact at mediqa-m3g 2024: Medical question answering in dermatology with multimodal learning. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Christine A. Sinsky, Tait D. Shanafelt, and Jonathan A. Ripp. 2024. The electronic health record inbox: Recommendations for relief. 37(15):4002–4003.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Jerrin John Thomas, Sushvin Marimuthu, and Parameswari Krishnamurthy. 2024. Ltrc-iiith at mediqa-m3g 2024: Efficient medical visual question answering with fine-tuned lightweight models. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Augustin Toma, Ronald Xie, Steven Palayew, Gary D. Bader, and BO WANG. 2024. Wanglab at mediqa-m3g 2024: Multimodal medical answer generation using large language models. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Parth Vashisht, Abhilasha Lodha, Mukta Maddipatla, Zonghai Yao, Avijit Mitra, Zhichao Yang, Junda Wang, Sunjae Kwon, and Hong Yu. 2024. Umassbionlp at mediqa-m3g 2024: Dermprompt - a systematic exploration of prompt engineering with gpt-4v for dermatological diagnosis. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. 2023. An early evaluation of gpt-4v(ision). *Preprint*, arXiv:2310.16534.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, and Meliha Yetisgen. 2024a. To err is human, how about medical large language models? comparing pretrained language models for medical assessment errors and reliability. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024b. Dermavqa: A multilingual visual question answering dataset for dermatology. *CoRR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

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