IKIM at MEDIQA-M3G 2024: Multilingual Visual Question-Answering for Dermatology through VLM Fine-tuning and LLM Translations

Marie Bauer[1] **Constantin Marc Seibold**[1] **Jens Kleesiek**[1,2,3,4] **Amin Dada**[1]
[1]Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), Essen, Germany
[2]Cancer Research Center Cologne Essen (CCCE), West German Cancer Center Essen
University Hospital Essen (AöR), Essen, Germany
[3]German Cancer Consortium (DKTK, Partner site Essen), Heidelberg, Germany
[4]Department of Physics, TU Dortmund, Dortmund, Germany

Abstract

This paper presents our solution to the MEDIQA-M3G Challenge at NAACL-ClinicalNLP 2024. We participated in all three languages, ranking first in Chinese and Spanish and third in English. Our approach utilizes LLaVA-med, an open-source, medical vision-language model (VLM) for visual question-answering in Chinese, and Mixtral-8x7B-instruct, a Large Language Model (LLM) for a subsequent translation into English and Spanish. In addition to our final method, we experiment with alternative approaches: Training three different models for each language instead of translating the results from one model, using different combinations and numbers of input images, and additional training on publicly available data that was not part of the original challenge training set.

1 Introduction

Over the past 25 years, various studies have discussed the shortage of dermatologists in the US (Kimball, 2003; Kimball and Resneck Jr, 2008; Ehrlich et al., 2017). At the same time, machine learning methods offer potential relief for the limited time available to dermatologists (Fogel and Kvedar, 2018) and, on some tasks, even exceed expert capabilities (Esteva et al., 2017). Recently introduced vision-language models (VLMs) showed promising capabilities in radiology and pathology visual question-answering (VQA) tasks (Moor et al., 2023; Wu et al., 2023; Thawkar et al., 2023; Liu et al., 2023; Chen et al., 2024). Therefore, it can be assumed that they also provide relief in the field of dermatology. However, there are no existing dermatology VQA datasets (Lin et al., 2023). Yet, VLMs need fine-tuning datasets to achieve the high accuracy required for medical tasks (Liu et al., 2023).

A possible data source for such tasks are telemedical records. Telemedicine describes triaging, diagnosing, and monitoring patients remotely through digital images and text messages (Waller and Stotler, 2018). Shortly after the outbreak of the COVID-19 pandemic, the availability of telemedicine services increased in parts of China (Hong et al., 2020; Song et al., 2020), providing new opportunities to create VQA datasets. Following these developments, the MediQA-M3G challenge (wai Yim et al., 2024a) is based on data from one of these telemedical platforms. The participants are offered photos of skin diseases and textual interactions between patients and medical professionals. While the original data is in Chinese, automated translations into English and Spanish were also provided. This raises several questions that we examined in the course of the challenge. First, there is the question of which model should be used on the Chinese data since all medical VLMs were trained in English. Another question is how helpful the training on the translated English and Spanish data is or whether problems such as translation errors and cultural differences are a hindrance.

To answer these questions, this paper compares various fine-tuning methods in preparation for our challenge submission. We first evaluate the usefulness of additional imaging data from two publicly available dermatological classification datasets in solving the challenge. We then compare multi-image training to training with a single image per data entry. Finally, we also test if training three different models for each language outperforms training a single model and translating its predictions into the other two target languages.

2 Related Work

2.1 VLMs

With the rapid development of LLMs (Hoffmann et al., 2022; Touvron et al., 2023a,b; Peng et al., 2023), various approaches have been pursued to extend these models to vision-language models

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 439–447
June 21, 2024

(VLMs) (Alayrac et al., 2022; Li et al., 2023b; Liu et al., 2023). This usually involves combining pre-trained LLMs and image models using dedicated architectures and training them on multimodal data. A notably straightforward yet effective architecture that has emerged from these efforts is LLaVA (Liu et al., 2023). Within this approach, a basic feed-forward network, comprising two layers is employed to map the image embeddings to the language embedding space of the LLM. Similarly to the development of specialized biomedical LLMs (Chen et al., 2023; Labrak et al., 2024; Xie et al., 2024), a modified version of LLaVA designed for biomedical applications, known as LLaVA-med (Li et al., 2023a), has been introduced. All of our solutions to this challenge are based on LLaVA-med.

2.2 Translation

Shortly after the release of ChatGPT and the subsequent focus on LLMs, their translation ability was explored (Hendy et al., 2023; Jiao et al., 2023; Bawden and Yvon, 2023). In contrast to previous neural machine translation (NMT) approaches that revolved around specialist language models trained on parallel translation corpora (Tiedemann and Thottingal, 2020; Costa-jussà et al., 2022), LLMs learn translation through vast pre-training and instruction tuning. Improvements over traditional NTM models include smoother translations (Hendy et al., 2023). However, these improvements are accompanied by higher translation error rates (Yao et al., 2024). An interesting observation by Hendy et al. (2023) is that GPT produces more accurate translations of noisy Chinese texts than traditional NMT models. Since the data in this challenge consists of Chinese consumer health questions, a translation with LLMs seems reasonable in this context. However, it also makes sense to evaluate a traditional NMT model due to the higher error rates of LLMs. Following its promising performance on medical downstream tasks (Dada et al., 2024), we used Mixtral-8x7B-Instruct (Jiang et al., 2024) for LLM-based translation of Chinese predictions and OPUS (Tiedemann and Thottingal, 2020) as the NMT model.

2.3 Consumer Health Question-answering

Previous works focused on consumer health question-answering but were only text-based (Ben Abacha et al., 2019; Ben Abacha and Demner-Fushman, 2019). Existing VQA datasets do not include consumer health inquiries and are based on radiology (Lau et al., 2018; Liu et al., 2021; Hu et al., 2023) and pathology images (He et al., 2020). Since no datasets are based on multimodal dermatology consumer health questions, there are currently no existing approaches for this task. Furthermore, using Chinese texts translated into English and Spanish is a novel approach that requires methods to address this setting adequately.

3 Challenge Dataset

The given dataset (wai Yim et al., 2024b) consists of clinical history and patient query examples. Along these textual inputs, one or multiple photos of the described skin disease were attached to the query. The gold labels consisted of one or multiple answers by Chinese dermatologists. All texts were machine-translated into English and Spanish without further information on which model was used for translation. One exception is the test set, which was translated manually. For the validation and test sets, annotators also provided a score indicating how complete an answer is concerning the query. Possible completeness scores were 0.0, 0.5, and 1.0, ranging from incomplete to entirely complete. As a metric, the deltaBLEU score (Galley et al., 2015) was computed between predictions and dermatologists' answers using the completeness score for weighting.

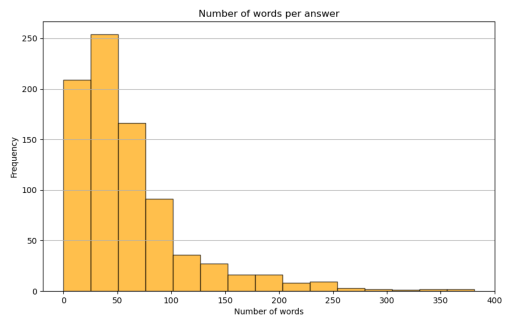

Figure 1: Histogram of number of words per dermatologist answer in English.

The training set consists of 842 patient queries with an average of 2.94 images per query. Additional 56 examples were provided as a validation set and 100 examples as a test set. Figure 1 shows the histogram of the number of words per dermatologist answer for the English training set. Most answers consist of only a few words, usually the diagnosis. However, some outliers are considerably longer, with over 315 words. These answers contain lengthy descriptions of the treatment and follow-up steps for the patient. While we manually analyzed the data, we could not find a consistent relationship between the type of request and the verbosity of the response.

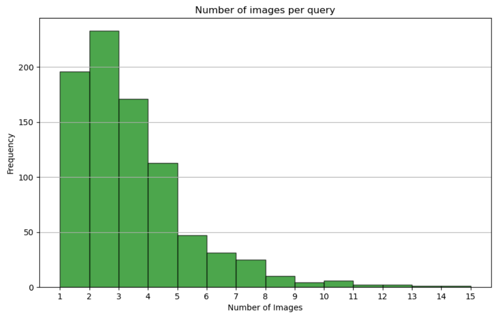

Figure 2: Histogram of number of images per patient query.

Figure 2 shows the histogram of the number of images per sample in the training set. Like the number of words per answer, queries usually have few images attached.

4 Methodology

The following section describes the different approaches for the challenge. We describe multi-image training, additional non-challenge data used, and our methods of translating LVM predictions into new target languages.

4.1 Training on multiple input images

The challenge data often provided multiple images for each input text (see Figure 2). This led to the question of whether all of them should be used together in a single text prompt, decreasing the number of training examples but potentially increasing the information available to the model in each case, or if each image should be used as input separately, thus increasing the number of training examples but potentially decreasing the quality of the input.

4.2 Additional fine-tuning data

In addition to the data presented by the challenge, we attempted to train the model on additional publicly available dermatological image datasets. For this, we employed Fitzpatrick17k (Groh et al., 2022), which contains approximately 17,000 labeled dermatological images and Dermnet[1], adding an additional 19,500 images. The aim was to increase the model's overall domain knowledge and to improve its performance in identifying common dermatological illnesses before training it on challenge data. We prompted the model to identify the illness in the picture using the image label as the prediction target.

4.3 Translation or language-specific fine-tuning

A central question for us was whether we should fine-tune three different models, one for each challenge language, or train a single model and translate the resulting predictions into the other two target languages. The first attempt had the potential to yield good results, especially in English, since LLaMA, which provided the base weights for LLaVA-Med, was only peripherally trained on Chinese and Spanish. On the other hand, the quality of training data was the highest in Chinese, since this was the language the data was sourced in, and translations were automatic and, in some places, inaccurate. This could lead to the model learning inaccurate terms, reflecting poorly in the test set because it was translated manually. In this case, translating the generated answers would be the preferred option. When translating with Mixtral, we prompted the model to generate an accurate translation of a Chinese forum post with medical content in Spanish and English, respectively. Figure 3 shows these prompts. To achieve higher-quality translations and to ensure the model would adhere to our instructions, we constructed 3 few-shot examples containing fictional example sentences that were similar in style but not originally contained in the training data. Finally, we post-processed with simple regex expressions to exclude additional remarks Mixtral often made, which were not part of the translation.

5 Results

Our best results were achieved by training LLaVA-med exclusively on Chinese challenge data, for only a single epoch, as more epochs to decrease performance. The learning rate was $2e-5$, with an overall batch size of 4 and 16 gradient accumulation steps. We did not make use of validation data in fine-tuning for our final submission. The resulting predictions were then translated into Spanish and English using Mixtral-8x7b-instruct. This method achieved a score of 7.05 BLEU for Chinese, 2.66 for English, and 1.36 for Spanish. (see Table 1). These represent the highest scores

[1]https://dermnet.com/

Figure 3: The system prompts used to generate translations from Chinese into English and Spanish

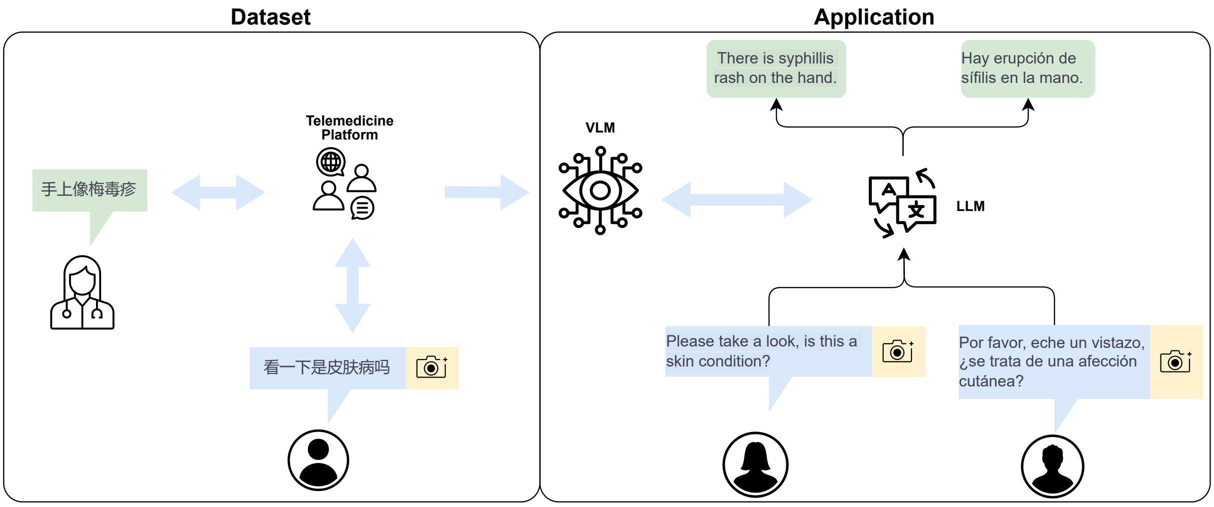

Figure 4: The left-hand side shows the dataset collection process. It consists of chat interactions between Chinese dermatologists and patients. Each patient inquiry contains a text and multiple photos of their skin disease. We train a VLM on the original Chinese examples. For the application of this model in other languages, we translate the model answers from Chinese to English and Spanish using an LLM.

achieved in the challenge for Chinese and Spanish. As mentioned in the previous section, this represents a fairly simple approach compared to other experiments we performed, which is visualised in its entirety in Figure 4.

6 Discussion

In addition to the main result described above, we performed several additional experiments with differing approaches, which in most cases led to worse performance than in the version we submitted. Table 1 gives an overview of these results. The following section gives some reasons for why additional training might have harmed model performance in this case and why a simple approach ended up achieving the highest scores.

6.1 Analysis of fine-tuning methods

It becomes apparent that additional datasets that were not originally part of the challenge worsen results by 0.61 points in the case of English and 1.08 points in the case of Spanish. Following up with fine-tuning on challenge data improved the score again slightly, but it does not come close to reaching the scores of training exclusively on challenge data. It is possible that this was due to the incompatibility of datasets, meaning that diagnoses contained in challenge data were not represented in Fitzpatrick or Dermnet. Additionally, challenge data often contained more complex tasks than correctly identifying what could be seen in the image, e.g., answering questions about potential treatments. Finally, the particular writing style of many entries in the challenge data and differing translations may also have played a role.

Mixtral-8x7b-instruct seems to outperform Opus as a translation option despite Opus being a group of models designed specifically for translation between set language pairs. One constraint expected to lead to Opus's poorer performance was that this model family only contains a model for Chinese to English and English to Spanish translations, but none for Chinese to Spanish, thus necessitating a translation first to English and then to Spanish. However, our results show this is not the case since the Spanish Opus translation outperforms the En-

Table 1: This table contains the various results we achieved with different fine-tuning methods. Datasets used: 1. M3G: original challenge data 2. DN: Dermnet 3. FP: Fitzpatrick17k

ID	Datasets	Training Language	Translation Method	Score (ZH / EN / ES)
1	M3G	Chinese	Mixtral	**7.05** / **2.66** / 1.36
2	M3G	English	-	- / 2.05 / -
3	M3G	Spanish	-	- / **1.58** / -
4	M3G	Chinese	Opus	7.05 / 0.60 / 0.99
5	FP	English	-	- / 0.47 / -
6	FP + M3G	English	-	- / 0.94 / -
7	DN	English	-	- / 0.57 / -
8	DN + M3G	English	-	- / 1.44 / -
9	DN + FP	English	-	- / 0.77 / -
10	DN + FP + M3G	English	-	- / 1.41 / -

glish translation.

The answer is unclear regarding whether one should train designated models for each language or translate results using a translator model. Translation outperforms designated training in the case of English but not in the case of Spanish.

In addition to the variables discussed up to this point, we trained LLaVA-Med by using multiple input images instead of a single one, which also worsened the results. Compared to our best result, training exclusively on Chinese challenge data, multi-image training only scored 0.63 BLEU. In general, increasing the number of input images during training seems to decrease LLaVA's performance. This might be because LLaVA models are usually only pre-trained using single images, so the model can not properly handle multi-image input. Another potential reason performance decreases with more input is the way multimodality is implemented in LLaVA models: image features are projected into the embedding space of the language model, thus effectively increasing input length. Longer contexts have been shown to decrease language model performance. (Levy et al., 2024)

Example

Prompt: Male, 16 years old. Got Pustule for 10 years at the hands and feet. Previously diagnosed as chronic eczema. Some of our colleagues suspected it is scabies. And hence was treated once as such, with no success. Later on, with a follow up on the medical history, it was revealed the lesion at the foot was the result of a scald. The lesions in the hands and feet have been there for 10 years. There was some relief after the use of anti-biotics. Now it mainly happens on the palms and side of the hands and feet. Once again, does it look like eczema? How to account for these small pustules? Picture uploaded before: The latest status:

LLaVA-med: Impetigo.

Reference 1: Impetigo.

Reference 2: Tinea pedis combined with fungal rash. Suggest to run a test on fungal infection.

Reference 3: What appear on the hand is eczema, impetigo. Exfoliative keratolysis at the feet

Figure 5: Example prediction by LLaVA-Med, translated into English using Mixtral.

6.2 Error analysis

Looking at the model predictions, it becomes clear that there are still several issues with its performance. Firstly, as the model mirrors challenge training data, it tends to gravitate towards very short and concise answers, simply stating a presumed diagnosis, see for example Figure 5.

These answers could, in some cases, agree with expert opinions but did not contain the same amount of additional information and did not directly answer the original question. There were also instances in which the model predicted a different but similar illness to the reference, often with a different descriptive adjective, see Figure 6.

Regarding translations, there were several differences between Mixtral and Opus. Opus tended to mistranslate medical terms. A possible reason is the added context that an LLM receives through the system prompt, clarifying that the input text has medical content. Additionally, it often failed to identify two connected Chinese characters as a single word, thus translating single characters liter-

Table 2: This table contains some examples that illustrate the differences in translation quality between Mixtral and Opus.

LLaVA-med prediction	Mixtral translation	Opus translation	Example ground truth
慢性湿疹	Chronic eczema.	Chronic wet rashes.	Could be chronic eczema.
季节性蜕皮症	Seasonal skin exfoliation.	Seasonal skinness.	Exfoliative keratolysis.
虫咬皮炎吧	Must be an insect bite dermatitis.	Worm bites skin inflammation.	Seborrheic dermatitis.
考虑毛囊炎	Considering folliculitis.	Think of cyborgitis.	[...] folliculitis or summer dermatitis [...]

Example

Prompt: See picture of the disease -maculopapular rash. The patient is male, 16 years old. Impression: appendicitis. Got influenza when in the hospital. After injection of antondine by the night shift physician, the skin in the affected area becomes itchy. They come to me for consultation the next day.

LLaVA-med: Skin rash disease?

Reference 1: Should be Measles type drug rash.

Reference 2: Drug rash.

Figure 6: Example prediction by LLaVA-Med, translated into English using Mixtral.

ally, missing the actual meaning. Table 2 contains a typical example for this: The Chinese word for 'eczema' consists of the characters for 'wet' and 'rash'. Opus interpreted these as distinct characters instead of a single word and thus reached an inaccurate translation. Opus also tended to choose general terms for some words instead of the correct scientific term. (E.g., simply 'inflammation' instead of 'dermatitis', see also Table 2) On the other hand, Mixtral achieved a relatively high quality of translations, given that it is neither officially trained on Chinese nor specifically biomedical data.

7 Limitations

The model we submitted has some limitations, excluding it from clinical use in its current state. Most importantly, even though our results scored the highest in two languages, the overall scores were very low. Significantly higher diagnosis accuracy must be achieved to make it useful in a clinical setting.

Secondly, due to the nature of the training data, the model often mimics the writing style of the forum posts it was trained on, leading to fewer professional expressions than expected in a clinical setting.

Similarly, since training data was obtained from Chinese sources containing frequent suggestions for using Traditional Chinese Medicine, the model made similar recommendations in some cases. This might not meet the standards of care in other countries. Thus, regional differences in care methods have to be considered when training similar models intended for clinical use in the future.

8 Conclusion

We present our submission to the Multilingual & Multimodal Medical Answer Generation task of the MediQA 2024 challenge. Our results compare well with other submitted approaches, but their quality is still insufficient for clinical use. This was partly because our method could not overcome hurdles presented by the challenge, such as short target predictions, translation issues, and regional differences in care methods. VLMs with better analytic capabilities in the medical domain must be created to achieve scores high enough for real-world applications. Nonetheless, the increased availability of telemedical records and the increasing availability of data from a variety of countries also presents an opportunity for medical LVM research.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Amin Dada, Marie Bauer, Amanda Butler Contreras, Osman Alperen Koraş, Constantin Marc Seibold, Kaleb E Smith, and Jens Kleesiek. 2024. Clue: A clinical language understanding evaluation for llms. *Preprint*, arXiv:2404.04067.

Alison Ehrlich, James Kostecki, and Helen Olkaba. 2017. Trends in dermatology practices and the implications for the workforce. *Journal of the American Academy of Dermatology*, 77(4):746–752.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

Alexander L Fogel and Joseph C Kvedar. 2018. Artificial intelligence powers digital medicine. *NPJ digital medicine*, 1(1):5.

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.

Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. 2022. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *arXiv preprint arXiv:2207.02942*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Zhen Hong, Nian Li, Dajiang Li, Junhua Li, Bing Li, Weixi Xiong, Lu Lu, Weimin Li, and Dong Zhou. 2020. Telemedicine during the covid-19 pandemic: Experiences from western china. *J Med Internet Res*, 22(5):e19577.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. 2023. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. KDD '23, page 4156–4165, New York, NY, USA. Association for Computing Machinery.

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.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Alexa B Kimball. 2003. Dermatology: a unique case of specialty workforce economics. *Journal of the American Academy of Dermatology*, 48(2):265–270.

Alexa Boer Kimball and Jack S Resneck Jr. 2008. The us dermatology workforce: a specialty remains in shortage. *Journal of the American Academy of Dermatology*, 59(5):741–745.

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.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *Preprint*, arXiv:2402.14848.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Xuan Song, Xinyan Liu, and Chunting Wang. 2020. The role of telemedicine during the covid-19 epidemic in china—experience from shandong province.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wen wai Yim, Asma Ben Abacha, Velvin Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024a. Overview of the mediqa-m3g 2024 shared task on multilingual and multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

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

Morgan Waller and Chad Stotler. 2018. Telemedicine: a primer. *Current allergy and asthma reports*, 18:1–9.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Huan He, Lucila Ohno-Machido, Yonghui Wu, Hua Xu, and Jiang Bian. 2024. Me llama: Foundation large language models for medical applications. *Preprint*, arXiv:2402.12749.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. Benchmarking llm-based machine translation on cultural awareness. *Preprint*, arXiv:2305.14328.

A Code availability

The code used to perform all experiments listed in this paper is available in this repository. [2]

[2] https://github.com/Shiniri/MediQA-M3G-Submission