NEUI at MEDIQA-M3G 2024: Medical VQA through consensus

R. Omar Chávez García[†*] and **Oscar Lithgow-Serrano**[†*]
[†] Dalle Molle Institute for Artificial Intelligence Research (IDSIA), USI-SUPSI, Switzerland
{omar.chavez, oscarwilliam.lithgow}@idsia.ch

Abstract

This document describes our solution to the MEDIQA-M3G: Multilingual & Multimodal Medical Answer Generation. To build our solution, we leveraged two pre-trained models, a Visual Language Model (VLM) and a Large Language Model (LLM). We fine-tuned both models using the MEDIQA-M3G and MEDIQA-CORR training datasets, respectively. In the first stage, the VLM provides singular responses for each pair of image & text inputs in a case. In the second stage, the LLM consolidates the VLM responses using it as context among the original text input. By changing the original English case content field in the context component of the second stage to the one in Spanish, we adapt the pipeline to generate submissions in English and Spanish. We performed an ablation study to explore the impact of the different models' capabilities, such as multimodality and reasoning, on the MEDIQA-M3G task. Our approach favored privacy and feasibility by adopting open-source and self-hosted small models and ranked 4th in English and 2nd in Spanish.

1 Introduction

Recent visual iterations of Large Language Models (LMM) explore a central concept that deals with multimodal inputs, known as visual instruction tuning. These studies result in sizable Visual Language Models (VLM) such as VisualBERT (Li et al., 2019), LLaVA (Haotian et al., 2023), MiniGPT-4 (Zhu et al., 2023) that demonstrate impressive results on natural instruction-following and visual reasoning capabilities.

The need for multimodal models is particularly pronounced in the medical domain. Medical Visual Question Answering (VQA) can assist in clinical decision-making, provide reliable and user-friendly answers to free-form questions, serve as a diagnostic tool, or act as a knowledgeable assistant, potentially alleviating the burden on the healthcare system and enhancing the efficiency of medical professionals. A mature and comprehensive medical VQA system could directly review patients' images and answer any questions, especially relevant when medical professionals may not be immediately available.

The MediQA-M3G task, which focuses on clinical dermatology multimodal query response generation, exemplifies this need. This task aims to automatically generate clinical responses given textual clinical history, user-generated images, and queries. The common challenges of VQA are amplified in the medical domain, where highly specialized knowledge must be leveraged in coordination with specific visual features from the images. This task is further complicated by the fact that the query, content, and images are provided by patients, which implies a highly heterogeneous text style, varying levels of description details, and, in the case of images, highly variable light, focus, zoom, and quality conditions.

We utilized a compact (1.86B parameters) Visual Language Model (VLM) named Moondream (Moondream AI, 2024) to evaluate the performance of small Language-Image Models (LIMs) on the M3G multimodal task. Moondream is built upon a Sigmoid loss for Language-Image Pre-training (SigLIP) and the Phi-1.5 language model. We fine-tuned the VLM using the provided training data, extending each case title and description to all the provided images.

The output of VLM might contain redundancies and short answers that deviate from the provided context in the query. We implemented a post-processing step of the VLM output to address this issue by constructing a new query for a fine-tuned BioMistral LLM. This step relies on the idea that we already have the context to improve the VLM answer. The context consists of the original query title and content from the test dataset cases and the

[*] All authors contributed equally.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 448–460
June 21, 2024

VLM response, which we refer to as image analysis. Along with the context, we used the general query *"What is the disease present in the photo? What is the treatment?"* We use the same pipeline for both English and Spanish submission entries.

2 Task definition

The *MEDIQA-M3G: Multilingual & Multimodal Medical Answer Generation* task focuses on the problem of clinical dermatology multimodal query response generation, a first of its kind, aiming to automatically generate clinical responses given textual clinical history, user-generated images, and queries (wai Yim et al., 2024a). This shared task is motivated and very in line with the rapid development of telecommunication technologies and the increased demand for remote clinical diagnosis and treatment. Unlike previous works focusing only on text or specific types of images, this task incorporates text and one or more images. Participants were given textual inputs, including clinical history and a query, along with associated images, and they were expected to generate a relevant textual response. The training data for this task was translated and adapted from Chinese datasets, and participants could opt to work in Chinese (Simplified), English, or Spanish for the test set (wai Yim et al., 2024b).

3 Related work

3.1 Large Language Models (LLM)

Integrating generative large language models (LLMs) has been pivotal in medical question-answering systems. Recent advancements have seen the adaptation of generalist LLMs like GPT-4 and Gemini to more specialized domains. However, the proprietary nature of such models limits their accessibility to research. This challenge has been addressed by the open-source movement, with models such as Llama 2 (Touvron et al.), Vicuna (Chiang et al.), and Mistral (Jiang et al.) providing a foundation for further innovations in medical LLMs. Multiple open-source LLMs based on decoder-only architecture have recently been developed for the medical domain, e.g., BioGPT (Luo et al.) and PMC-LLaMA (Wu et al.). Two notable recent contributions in this space are MediTron (Chen et al.) and BioMistral (Labrak et al.). MediTron, leveraging Llama-2, has been pre-trained on a vast corpus of medical literature to offer medical reasoning capabilities. In parallel, BioMistral adapts the Mistral model to the biomedical domain, showing the potential of merging techniques (Yu et al.) on pre-trained models to enhance performance and out-of-domain generalization. In particular, BioMistral, through techniques like DARE, has demonstrated improved robustness in multilingual settings, a key factor in real-case global medical applications.

The massive increase in the size of large language models and, by extension, visual language models to hundreds of billions of parameters has unlocked various emerging capabilities that have redefined the landscape of natural language processing and a plethora of downstream tasks. A common challenge remains whether such emergent abilities can be achieved at a smaller scale using strategic choices for training, e.g., data selection. Proposals such as the Phi family models aim to answer this question by training small language models (SLMs) that achieve performance on par with models of much larger (yet still far from the frontier models) (Javaheripi and Bubeck, 2024). Their success relies upon training data quality and the scalability of their smaller models.

3.2 Multimodality

The recent progress of multimodal models in the medical domain is highlighted by the progress in large vision language models such as Flamingo (Alayrac et al.), GPT-4V, and Gemini (Gemini Team et al.), which have demonstrated remarkable capabilities in executing instructions, engaging in dialogues, and managing image-based tasks. Such advancements show the potential of fusing vision encoders with large language models (LLMs) to create systems that can interpret and respond to complex queries involving textual and visual inputs. However, increased hardware demands, longer test times, slower inference speeds, and privacy concerns when used as cloud services are challenges to their use in real-case scenarios, especially for the case of medical applications.

End-to-end Vision-Language Pre-training. End-to-end vision-language pre-training (VLP) has been used to develop multimodal foundation models that excel in various vision-and-language tasks. Despite the effectiveness of these models, thanks to the evolution of architectures, learning objectives, and strategies such as contrastive learning and image-text matching, their use is hindered by requiring substantial computational resources for end-to-end training in large image-text pair datasets. Another limitation is the lack of leverage in existing unimodal pre-trained models. (Faria et al.; Lin et al.)

Modular Vision-Language Pre-training. In contrast, this approach involves modular VLP methods that utilize off-the-shelf pre-trained models, keeping them frozen during VLP. This includes techniques that freeze the image encoder, leveraging pre-trained models like CLIP (Radford et al.), and methods that freeze the language model to harness the knowledge from LLMs for vision-to-language tasks. A challenge in this approach is aligning visual features with the text space. BLIP-2 (Li et al.) is a successful recent approach that efficiently uses frozen image encoders and LLMs for various vision-language tasks with reduced computational costs.

Multimodal Instruction-following Agents. Instruction tuning has been crucial in reducing complexity and costs by training the model to handle various tasks represented by different instructions, thus eliminating the need for separate models for each application. Common architectures for instruction-following Large Language Models (LLMs) include a pre-trained visual backbone, a pre-trained LLM, and a vision-language cross-modal connector. Notable recent implementations include BLIP-2 (Li et al.) and LLaVA (Liu et al., 2023b,a) models. These represent significant steps in leveraging pre-trained models and visual instruction-tuning to enhance the capabilities of multimodal systems. The introduction of LLaVA-Phi (Zhu et al.) further exemplifies the trend toward creating efficient and compact models capable of delivering high performance in real-time applications. These developments point to AI systems' growing capabilities in understanding and acting upon instructions involving both visual and textual information.

Medical Visual Question Answering. Medical VQA can potentially transform clinical decision-making and patient engagement (Lin et al.). The unique challenges of the medical domain, such as privacy requirements, the need for expert annotation, and the integration of medical knowledge bases, are part of the complexity of developing effective Medical VQA systems. Dataset quality and diversity are among the most impactful limitations that must be addressed to advance the field. Although the LLMs and LMMs are adapted to the medical domain and already trained for instruction-following, it is often observed that their zero-shot and few-shot performance can be further enhanced by performing a complementary, focused SFT stage on specific tasks. Notably, task-specific models trained on carefully curated datasets have frequently outperformed generalist models of similar size, especially in highly specialized domains such as medicine.

4 Methodology

We utilized a compact (1.83B parameters) Visual Language Model (VLM) named Moondream (Moondream AI, 2024) to evaluate the performance of small Language-Image Models (LIMs) on the M3G multimodal task. Moondream is built upon a Sigmoid loss for Language-Image Pre-training (SigLIP) (Beyer et al., 2022) and the Phi-1.5 language model, a Transformer with 1.3B parameters (Li et al., 2023; Microsoft Research, 2023). In such a setup, a contrastively pre-trained model provides significantly more useful tokens than one classification pre-trained model (Zhai et al., 2023). Figure 1 shows the schematic of the proposed solution involving the VLM and the BioMistral-7B-DARE (Labrak et al.) LLM as a specialized stage for final response consolidation.

4.1 Training

Fine-tuning VLM. We fine-tuned the VLM using the whole provided training data, extending each case title and description to all the provided images (see Table 1). We employed the flash attention algorithm to mitigate memory issues during training and inference. Our hardware setup was limited to a single NVIDIA RTX 3090 GPU for fine-tuning and inference.

The motivation behind this training dataset is to increase the number of training samples, given the reduced number of clinical cases in the provided training data. The caveat of this approach is that although we consider each augmented sample as valid, there might be responses that overlap, complement, or contradict a valid clinical response.

Fine-tuning LLM. Our team, having participated in the MEDIQA-CORR (Ben Abacha et al., 2024a) task, leveraged the LLM fine-tuned for that task. Specifically, we instruction-tuned the BioMistral-7B-DARE on the MEDIQA-CORR dataset (Ben Abacha et al., 2024b). For this, we

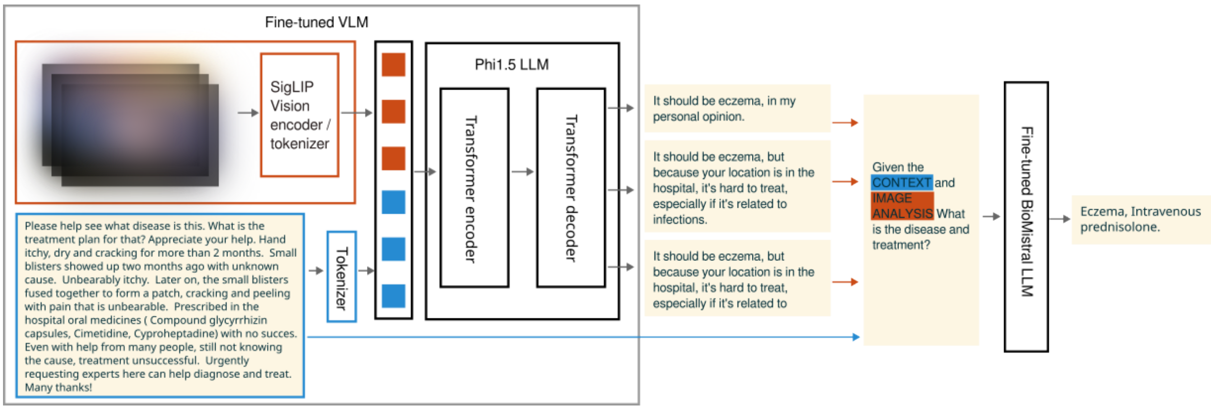

Figure 1: Overview of the proposed solution[1]. The contrastively pre-trained SigLIP vision model encodes the image into visual tokens individually. These visual tokens are passed along with a query to the Phi 1.5 LLM, producing responses for individual images. A consolidation response stage is performed via the fine-tuned Biomistral LLM using the VLM responses and context from the original query.

mapped the labeled dataset into three types of instructions: classifying if the statement had an error or not, detecting the culprit sentence, and correcting a given erroneous sentence to ensure consistency with the rest of the clinical text. The Supervised Fine-Tuning (SFT) was performed using the parameter-efficient method LORA on an NVIDIA A100-80G for four epochs. Without further training, we then used this MEDIQA-CORR fine-tuned model in the M3G task.

4.2 Inference strategies

Strategy-1: Direct inference with VLM. We constructed the output by performing inference on each image of each case in the test dataset. This step means that for one case, we request the fine-tuned VLM with our query and each of the case's images. Finally, all VLM responses for a case, as many as images in the test case, were concatenated as the final response (see Table 2:2). The results of this strategy outperform the baseline obtained using the non-fine-tuned VLM (see Table 2:1).

Strategy-2: Two-stage inference (VLM + LLM) The output of the previous approach might contain redundancies and short answers that deviate from the provided context in the query. To address this issue and to harness knowledge from a bigger specialized model, we implemented a two-stage strategy that augmented the previously described *Direct inference* strategy with a post-processing step. This step relies on the idea of leveraging the arguably better reasoning capabilities of a bigger specialized LLM to better harness the provided case information, i.e., query title and content, along with the VLM answers to generate a final response. Specifically, we requested the LLM with a prompt consisting of the query: "What is the disease present in the photo? What is the treatment?"; the context: dataset query title and content; and the image analysis: list of VLM responses (see Table 2:3). Table 3 shows examples of the composite input of this step and the resulting consolidated response.

Regarding multilingualism, Strategy-1 was built considering only one language data stream, English. The VLM was fine-tuned using only the English queries, content, and responses. However, as the LLM we employed in Strategy-2 has multilingual capabilities (see sec. 3.1), we also applied the post-processing step of Strategy-2 to the Spanish version of the cases. We provided a prompt with the query and context in Spanish but with the English image analysis. We added additional prompt instructions to the model to request Spanish responses exclusively. As a result of this change, we could provide output for the Spanish version of the task (see Table 2:3).

5 Results and analysis

Results during competition. From the official results during the competition (Table 2 ids: 1-3), we observe that by fine-tuning the VLM (id: 2), we obtained a significant improvement, with a `deltableu` of 0.595, which is more than a two-fold enhancement over the baseline non-fine-tuned version (id: 1) that held a `deltableu` of 0.231. Furthermore, the implementation of Strategy-2 (id: 3) marked a substantial leap, exhibiting a quadruple

[1]MEDIQA-M3G dataset contains images of medical conditions that may be sensitive and/or graphic in nature.

Original sample (single language)	Training sample
case: `ENC00018`	**sample:** `ENC00018_image1_response1`
(image1, image2)	(image1)
from: human; **value:** (title) `View image` (content) `Female, 19 years old, has had a hard lump in her ear for three months, as hard as a wooden board, with no sense of fluctuation. After incision, a white dense substance was found. What kind of cyst could this be?`	**from:** human; **value:** (title) `View image` (content) `Female, 19 years old, has had a hard lump in her ear for three months, as hard as a wooden board, with no sense of fluctuation. After incision, a white dense substance was found. What kind of cyst could this be?` (augmented query) `What is the disease in the photo? What is the best treatment?`
from: response 1; **value:** `Erythema annulare centrifugum? Use licorice decoction with corticosteroid ointment.`	**from:** agent ; **value:** `Erythema annulare centrifugum? Use licorice decoction with corticosteroid ointment.`
from: response 2; **value:** `I think it still looks like urticaria, continue with the anti-allergy treatment.`	
from: response 3; **value:** `I think the likelihood of urticaria is the highest, but the skin lesions at the root of the thigh are hard to explain, so erythema annulare cannot be ruled out either...`	

Table 1: Training example used for fine-tuning the VLM. We augment the training query (represented by the title and content case) with the standard query from the challenge description. We generate a training sample per each image and response combination. Hence, each case in the training dataset will generate $I \times R$ training samples, where I is the number of images in the case, and R is the number of responses for the selected language.

increase in performance for the English language tasks, as indicated by a `deltableu` of 2.133 compared to the 0.595 achieved by Strategy-1 (id: 2). When applied to Spanish, Strategy-2 (id: 3) show a significant drop but still got a competitive performance with a `deltableu` of 0.974. Our best runs (id: 3) were placed at the 4th and 2nd positions for English and Spanish, respectively.

5.1 Ablation study

We conducted an ablation study to assess the impact of various components in our best strategy (Strategy-2). We can isolate and understand their contributions to the strategy's effectiveness by systematically removing or altering specific model elements. Our analysis focuses on three primary objectives: investigating the Unimodal Bias phenomenon, assessing the extra reasoning capacity contribution of the Large Language Model (LLM), and evaluating the impact of training the LLM on a specialized dataset for error detection and correction in clinical notes.

Investigating the Unimodal Bias Phenomenon. To explore the Unimodal Bias and the impact of incorporating visual modality, we performed experiments 4 and 5 (see Table 2). We follow the same pipeline as in Fig. 1 without using the input images for the unimodal experiments. Thus, the VLM only sees the test case's title and content text inputs as prompts, i.e., we remove the references to "image" from the prompt. In experiment id:4 our strategy involved employing the Visual Language Model (VLM) without providing visual inputs, relying solely on textual content. This setup mimics Strategy-1 but aims to quantify the absence of visual modality. Experiment id:5 followed a similar approach, utilizing both the VLM and LLM without visual inputs, akin to Strategy-2. As seen in Table 2, the results – `deltableu` scores of 1.418 and 0.968 for English and Spanish, respectively in id: 4 and 0.328 for English in id: 5– indicate the positive impact of using both modalities in this task. The decrements in `BERTscore` and `deltableu` metrics suggest that relevant information in the encoded & tokenized image input is helping, along with textual case input, to determine the test case queries.

Assessing the Extra Reasoning Capacity of the LLM. The comparison of Strategy-2's performance under varying conditions—specifically when the LLM is provided with both the case context and VLM responses versus when it only receives the VLM responses for summarization—sheds light on the LLM's reasoning ability. Experiment id:6 explores this, allowing us to distill the LLM's added value in synthesizing and

	id	Strategy	EN	ES
Test	1	Moondream	0.231	-
	2	Moondream-FT	0.595	-
	3	Moondream-FT + BioMistral-FT	**2.133**	0.974
Test_after	4	Moondream-FT :: w/o visual	0.328	-
	5	Moondream-FT + BioMistral-FT :: w/o visual	1.418	0.968
	6	Moondream-FT + BioMistral-FT :: w/o context	1.183	-
	7	Moondream-FT + BioMistral :: w/o FT-LLM	1.963	**1.745**

Table 2: Official scores (`deltableu`) of the different submitted strategies for English (EN) and Spanish (ES). Stages, Test: during competition, Test_after: after the end of competition. The best scores by language appear in bold.

reasoning over the provided information. With a `deltablue` score of 1.183 in English, this experiment shows how much the LLM's reasoning capabilities, beyond mere summarization, contribute to generating more correct and contextually aware responses.

Evaluating the LLM's Training on Error Detection and Correction. BioMistral LLM utilizes Mistral as its foundation model. It is further pretrained on PubMed Central (a dataset containing citations and abstracts of biomedical literature), making it a top performer in medical question-answering benchmarks in English. Experiment id:7 investigates the relevance of the ability of error detection and correction within clinical notes by exploring the original BioMistral against one fine-tuned on the CORR dataset. This experiment examines the hypothesis that an LLM trained for error detection&correction could better integrate VLM responses with the textual case content, especially in correcting inconsistencies in VLM responses ("diagnostic"). The results from this experiment, 1.963 for English and 1.745 for Spanish, demonstrate the potential benefits of specialized fine-tuning for tasks out of the LLM's immediate domain expertise, highlighting the enhanced capability for error correction and the generation of coherent and accurate clinical responses.

5.2 Discussion

Analyzing the results of the competition phase and of the ablation study (see Figure 2 & Table 2), we observe that when using the non-fine-tuned version of BioMistral, we obtain the smallest drop in performance, a mere 7%. In contrast, a more significant drop in scores, a 33% degradation, was observed when the visual input was removed. Interestingly, the loss was even higher, at 44%, when neglecting the reasoning capabilities of the LLM. This suggests that the analysis and synthesis, i.e., **consensus generation** capacity of the LLM, is a key component of the strategy.

All the ablation experiments in Table 2, except for the experiment id:7, for Spanish, resulted in lower scores. Interestingly, the Spanish version in experiment id:7 scored the highest and surpassed any published run in the leaderboard to the best of our understanding. We hypothesize that by fine-tuning BioMistral on the CORR dataset (which is only in English), we not only steered the LLM towards a narrow set of tasks, specifically clinical error detection and correction but also disrupted the model's capacity to handle other languages due to the monolingual nature of the training set and the prevalent use of English in the pretraining corpora. This leads to an intriguing inquiry: how robust is the multimodal capacity attributable to merging methods like DARE (Yu et al., 2024) when subjected to monolingual posterior fine-tunings? This question warrants further investigation. Furthermore, removing the visual input has almost no impact on the performance in the Spanish version (id:3 vs. id:5). We hypothesize that this is also the result of our BioMistral-FT version's degraded multilingual capacity, which makes it unable to integrate the case context in Spanish with the VLM responses in English.

The competition results and ablation analysis clearly indicate that in Strategy-2, all components work collaboratively for the better. Even the fine-tuned version of BioMistral, which had the lowest impact, positively contributed to the final score. This shows the key role of consensus generation.

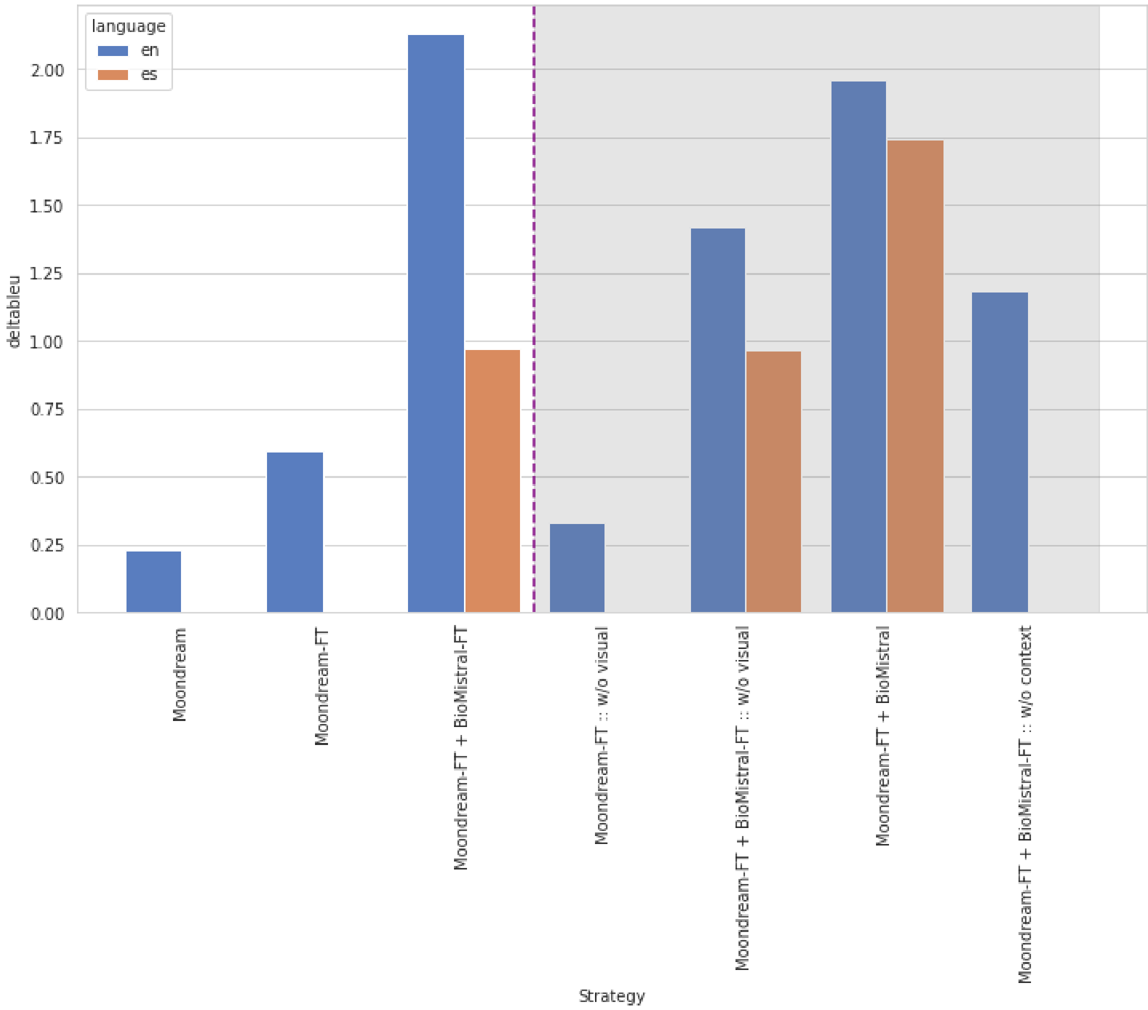

Figure 2: Deltableu scores of our submissions. The strategy used is represented on the X-axis. Scores for English are in blue and for Spanish in orange. The shaded area represents the submissions in the after-test stage.

By integrating multiple independent responses from the multimodal model and re-analyzing the case context, the strategy generates a revised final response, which is more contextually accurate.

LLMs, and by extension VLMs, differ significantly from prior deep learning methods regarding their scale, capabilities, and broad potential impact. For instance, these models are trained on massive datasets and use billions of parameters, resulting in considerable complexity. Models of this scale require significant hardware resources for training, fine-tuning, and, some, even inference. Privacy in medical applications of LLMs is paramount, and the possibility of training and testing this kind of model on-site is critical. Relying on third-party hardware providers to store or process medical data becomes a privacy risk (Khullar et al., 2024; Meskó and Topol, 2023). Our proposed pipeline considers this requirement when dealing with the delicate nature of the images in the training dataset. To do so, we explored the use of compact Visual Language Models and their performance in the M3G task. Our results provided a promising perspective, even with the limited data we utilized for training and the conservative score we obtained in the challenge.

6 Limitations

Our proposed approach holds significant promise for the VQA problem in clinical dermatology , but several limitations associated with deploying VLMs and LLMs in real-world medical settings necessitate careful consideration.

We are optimistic about the potential of our 2-step method, which is designed to consolidate multiple responses from image-text query pair analysis into a single, consensus-based solution. This approach allows us to utilize simpler VLMs, initially intended for single-image scenarios, in settings with an unknown number of multiple images. However, we acknowledge that this flexibility comes with the cost of posing as many queries to the VLM as there are images in a single case. This could increase computational costs, potentially making the solution computationally impractical for real-world deployment.

Another limitation is that VLMs, which were aligned with domain-specific images and texts during pretraining, are observed better to leverage domain-specific training examples during the instruction tuning phase. However, in our setting, the VLM lacks alignment for visual-medical texts and relies solely on fine-tuning to generate the most appropriate answers. This lack of alignment makes the VLM more demanding for instruction tuning data.

Another crucial problem with VLMs based on the pre-trained vision encoder is resolution. They are trained and also expect to analyze the full image input. However, for some specific cases, and even if the input is big enough, the focus of the query relies on certain zones of the image –in extreme cases, these zones are tiny compared to the image resolution. The M3G dataset showcases this very problem. Most dermatology-related images in either training, validation, or test datasets contain wide shots of the patient's limbs, and only a tiny region of the image provides valuable visual cues. We envisioned exploring techniques such as Visual Search (Wu and Xie, 2023) and Visual Cropping (Zhang et al., 2024) that can help us tackle this issue without compromising the size of our affordable VLM.

Regarding the LLM component, even if we are within the considered "small" scale, the computational demands of these models are substantial. Operating such models requires significant computational resources, which may not be feasible in all clinical environments. This issue can hinder our proposed solution's scalability and practical deployment in resource-limited settings. Moreover, LLMs are prone to generating "hallucinations" or outputs that may include incorrect or misleading information. This phenomenon is particularly concerning in the medical field, where accuracy is crucial to avoid misdiagnoses or inappropriate treatments. Intrinsic hallucinations, where outputs logically contradict known facts, and extrinsic hallucinations, where outputs cannot be verified, both pose serious risks in clinical applications.

Additionally, data bias and patient privacy are critical. LLMs trained on biased data can perpetuate or amplify these biases, leading to skewed or unfair medical advice. For example, the competition dataset observed a frequent recommendation based on traditional Chinese medicine. Thus, a model trained on this dataset may exhibit a predisposition, favoring a certain type of recommendation, irrespective of local or user-specific preferences. Given the sensitive nature of medical data, ensuring patient privacy while using such models is also paramount. Furthermore, updating these models with new medical knowledge remains a complex and resource-intensive process. This is problematic in the fast-evolving field of medicine, where staying current with the latest research and clinical findings is essential. For example, if a new Adverse Drug Effect is discovered, it is vital to update the models' knowledge promptly.

Finally, it is necessary to be aware that although this shared task is a crucial step towards better understanding and addressing the relevant task of automatically generating clinical responses given the textual clinical history and user-generated images, similar to existing benchmarks and metrics, it does not imply a comprehensive assessment of the performance of the system in real-medical contexts. Metrics such as trustworthiness, helpfulness, explainability, and faithfulness are crucial for clinical applications, and addressing these issues involves not only technical advancements in the architecture and training of LLMs but also close collaboration with medical professionals to ensure the clinical validity and ethical deployment of these technologies.

In conclusion, while our solution shows promise in addressing the MEDIQA-M3G task, limitations must be addressed to make it suitable for clinical use. Further exploration of optimization strategies, evaluation with other metrics, and collaboration with medical professionals are necessary to improve our approach's clinical relevance and effectiveness in real-world healthcare settings.

7 Future directions

We plan to incorporate a broader array of medical and health-related datasets into our training regimen to enhance our models' domain-specific accuracy and relevance. Specifically, we aim to utilize the Skin Condition Image Network (SCIN) dataset (Ward et al., 2024) focused on dermatology, including structured and unstructured textual data.

Moreover, we are interested in exploring the potential benefits of integrating data from various clinical specialties into our training process to see how this affects the model's performance and applicability across different medical fields.

We are particularly keen on incorporating retrieval-augmented generation (RAG) strategies related to the challenges of model knowledge updating and mitigating hallucinations. These strategies leverage existing related medical knowledge during the inference phase to enhance the factuality of the generated responses. By doing so, we expect to improve the reliability and accuracy of the model outputs, which is crucial for clinical applications.

Finally, we recognize the importance of interdisciplinary collaboration in developing medical VLMs and LLMs. Therefore, we are already in plans to initiate partnerships with medical professionals who can provide valuable insights, contribute relevant training data, and help define the desired outcomes for these technologies. Their involvement is critical not only in the development phase but also in testing these models in real-world clinical scenarios to ensure they meet both practical clinical needs and high standards of medical care.

8 Conclusions

We explored a solution to the clinical dermatology multimodal query response generation task and proposed a pipeline that can be expanded to similar multimodal tasks. We leverage performant pre-trained language models, fine-tuning the small VLM to adapt to the clinical task. We also show how the pipeline adapts to the multilingual complementary problem by relying on the multilingual capabilities of the pre-trained LLM. Participating in this challenge represented a feasibility study and opened several work perspectives for multimodal medical applications.

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 L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A Visual Language Model for Few-Shot Learning. 35:23716–23736.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of l1-regularized log-linear models. In *International Conference on Machine Learning*.

Asma Ben Abacha, Wen wai Yim, Yujuan Fu, Zhaoyi Sun, Fei Xia, and Meliha Yetisgen. 2024a. Overview of the mediqa-corr 2024 shared task on medical error detection and correction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Asma Ben Abacha, Wen wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2024b. Medec: A benchmark for medical error detection and correction in clinical notes. *CoRR*.

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. 2022. Big vision.

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. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *Preprint*, arxiv:2311.16079.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.

Ana Cláudia Akemi Matsuki Faria, Felype de Castro Bastos, José Victor Nogueira Alves Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves Santos. Visual Question Answering: A Survey on Techniques and Common Trends in Recent Literature. *Preprint*, arxiv:2305.11033.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, and Others. Gemini: A Family of Highly Capable Multimodal Models. *Preprint*, arxiv:2312.11805.

Liu Haotian, Li Chunyuan, Wu Qingyang, and Jae Lee Yong. 2023. Visual instruction tuning. In *NeurIPS23*, volume 6.

Mojan Javaheripi and Sébastien Bubeck. 2024. Phi-2: The surprising power of small language models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, de las Casas, Diego, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *Preprint*, arxiv:2310.06825.

D. Khullar, X. Wang, and F. Wang. 2024. Large language models in health care: Charting a path toward accurate, explainable, and secure ai. *J GEN INTERN MED*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. *Preprint*, arxiv:2402.10373.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *Preprint*, arxiv:2301.12597.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical Visual Question Answering: A Survey. 143:102611.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *Preprint*, arxiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arxiv:2304.08485.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie Yan Liu. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. 23(6):1–12.

B. Meskó and E. J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ Digit Med*.

Microsoft Research. 2023. The language model phi-1.5.

Moondream AI. 2024. moondream: a computer-vision model can answer real-world questions about images. Visited, March 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *Preprint*, arxiv:2103.00020.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Hugo Touvron, Thibaut Lavril, Xavier Martinet, Marie-anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Guillaume Lample, and Meta Ai. LLaMA : Open and Efficient Foundation Language Models.

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

Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep Kumar S, Tiya Tiyasirichokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, Christopher Semturs, and Pooja Rao. 2024. Crowdsourcing dermatology images with google search ads: Creating a real-world skin condition dataset. *Preprint*, arXiv:2402.18545.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. *Preprint*, arxiv:2304.14454.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *Preprint*, arXiv:2312.14135.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language Models are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. *Preprint*, arxiv:2311.03099.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *Preprint*, arXiv:2311.03099.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. 2024. Towards perceiving small visual details in zero-shot visual question answering with multimodal llms. *Preprint*, arXiv:2310.16033.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. LLaVA-Phi: Efficient Multi-Modal Assistant with Small Language Model. *Preprint*, arxiv:2401.02330.

A Appendix

A.1 Strategy-2

A.2 Submission results

-	Content
Case	**ENC00932**

Prompt	`### Instruction: Given the CONTEXT and IMAGE ANALYSIS What is the disease and treatment?`
	
	`### Input:`
	`CONTEXT: Please help take a look Is this onychomycosis? The patient is female, 60 years old. Got onychomycosis. Use Levothyroxine Sodium tablet daily. The nail is thin in the front, but rather thick in the rear part.`
	
	`IMAGE ANALYSIS: Onychomycosis, nail detachment? Onychomycosis, nail detachment? Onychomycosis, nail detachment?`

Response	`Onychomycosis, nail detachment?`
Official responses	`- It doesn't look like onychomycosis.` `- Should be onychomycosis.`
Case	**ENC00952**
Prompt	`### Instruction: Given the CONTEXT and IMAGE ANALYSIS What is the disease and treatment?` `### Input:` `CONTEXT: Please help see what disease is this. What is the treatment plan for that? Appreciate your help. Hand itchy, dry and cracking for more than 2 months. Small blisters showed up two months ago with unknown cause. Unbearably itchy. Later on, the small blisters fused together to form a patch, cracking and peeling with pain that is unbearable. Prescribed in the hospital oral medicines (Compound glycyrrhizin capsules, Cimetidine, Cyproheptadine) with no succes. Even with help from many people, still not knowing the cause, treatment unsuccessful. Urgently requesting experts here can help diagnose and treat. Many thanks!`  `IMAGE ANALYSIS: It should be eczema, in my personal opinion. It should be eczema, but because your location is in the hospital, it's hard to treat, especially if it's related to infections. It should be eczema, but because your location is in the hospital, it's hard to treat, especially if it's related to infections.`
Response	`Eczema, Intravenous prednisolone`
Official responses	`- Chapped eczema` `- Chronic eczema. Take special care on contact with specific material or work related.`
Case	**ENC00925**

Prompt	`### Instruction: Given the CONTEXT and IMAGE ANALYSIS What is the disease and treatment?`
	
	### Input:
	`CONTEXT: Papules and lumps at the lower left side of the abdomen for 7 months. The infant patient is 10 months old, and got the disease for 7 months. Got white flat papules and plaque on lower left abdomen. (see pictures) There is no pathological examination yet. Would experts here help diagnose.`
	
	`IMAGE ANALYSIS: Consider whether there is a lesion under the skin, the specific direction is difficult to consider.`

Response	`Langerhans cell histiocytosis`
Official responses	`- Could be Verrucous nevus` `- Verrucous nevus`

Table 3: Examples of prompts given to the LLM for the Strategy-2. Note: In this table, prompts are formatted in the Alpaca style with added line breaks for improved readability.

run_id	lang	stage	shortdesc	deltaBLEU	BERTScore
52859	en	test	Moondream	0.231	0.810
52872	en	test	Moondream-FT	0.595	0.851
52897	en	test	Moondream-FT + BioMistral-FT	2.133	0.850
54076	en	test_after	Moondream-FT :: w/o visual	0.328	0.842
54086	en	test_after	Moondream-FT + BioMistral-FT :: w/o visual	1.418	0.846
54091	en	test_after	Moondream-FT + BioMistral	1.963	0.829
54092	en	test_after	Moondream-FT + BioMistral-FT :: w/o context	1.183	0.860
52899	es	test	Moondream-FT + BioMistral-FT	0.974	0.814
52908	es	test	Moondream-FT + BioMistral-FT	0.974	0.814
54085	es	test_after	Moondream-FT + BioMistral-FT :: w/o visual	0.968	0.810
54173	es	test_after	Moondream-FT + BioMistral	1.745	0.809

Table 4: All team submissions by language and in chronological order.