WangLab at MEDIQA-M3G 2024: Multimodal Medical Answer Generation using Large Language Models

Ronald Xie[1,3] **Steven Palayew**[1,3] **Augustin Toma**[1,3]
Gary D. Bader[1,4,5] **Bo Wang**[1,2,3,6]

[1]University of Toronto [2]Peter Munk Cardiac Centre, University Health Network
[3]Vector Institute [4]Princess Margaret Cancer Centre, University Health Network
[5]Lunenfeld-Tanenbaum Research Institute, Sinai Health System
[6]AI Hub, University Health Network

{augustin.toma, ronald.xie, steven.palayew}@mail.utoronto.ca, gary.bader@utoronto.ca, bowang@vectorinstitute.ai

Abstract

This paper outlines our submission to the MEDIQA2024 Multilingual and Multimodal Medical Answer Generation (M3G) shared task. We report results for two solutions under the English category of the task, the first involving two consecutive API calls to the Claude 3 Opus API and the second involving training an image-disease label joint embedding in the style of CLIP for image classification. These two solutions scored 1st and 2nd place respectively on the competition leaderboard, substantially outperforming the next best solution. Additionally, we discuss insights gained from post-competition experiments. While the performance of these two solutions have significant room for improvement due to the difficulty of the shared task and the challenging nature of medical visual question answering in general, we identify the multi-stage LLM approach and the CLIP image classification approach as promising avenues for further investigation.

1 Introduction

An increased demand for healthcare services and recent pandemic needs have accelerated the adoption of telehealth, which was previously underused and understudied (Shaver, 2022; wai Yim et al., 2024a). There has been significant recent interest in integrating artificial intelligence (AI) into telehealth Ma et al., 2024; Toma et al., 2023, as these technologies have the potential to enhance and expand its ability to address important healthcare needs (Sharma et al., 2023). The task of consumer health question answering, an important part of telehealth, has been explored actively in research. However, the focus of this existing research has been on text (Ben Abacha et al., 2019), which is limiting as medicine is inherently multimodal in nature, requiring clinicians to work not just with text but also with imaging among other modalities (Corrado and Matias, 2023).

To help address this gap, the MEDIQA-M3G shared task was proposed (wai Yim et al., 2024a). This task requires the automatic generation of clinical responses given relevant user generated text and images as input, with a specific focus on clinical dermatology (wai Yim et al., 2024a).

This work describes our submission to this task. We explored two standalone solutions, one involving two consecutive API calls to the recently released Claude 3 Opus model (Anthropic) and the other trains a joint image-disease label embedding model using CLIP (Radford et al., 2021) for image classification. These two strategies took 1st and 2nd place respectively during the competition. While our strategy's effectiveness relative to other submissions highlight that Claude 3 Opus and multi-stage LLM frameworks have potential value in the area of multi-modal medical AI, both our solutions' performance is limited despite their leaderboard success, highlighting the difficulty of the shared task and the unsolved challenge of medical visual question answering.

2 Shared task and provided dataset

The MEDIQA-M3G competition focuses on the problem of clinical dermatology multimodal query response generation. The inputs include text which give clinical context and queries, as well as one or more images associated with the case (wai Yim et al., 2024b). The task is to generate responses to these cases resembling those made by medical professionals in the field of dermatology. Participants have the option to generate these responses in three languages: Chinese (Simplified), English, and Spanish. (wai Yim et al., 2024a)

The dataset consists of 842 train, 56 validation, and 100 test cases. Each case consists of one or more images of skin conditions, their accompanying query text which may or may not include clinical context, patient queries, additional details

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 624–634

regarding the disease and in some cases possible diagnosis. Finally, for each case there are multiple responses made by one or more medical professionals, which are used as targets to score the model predictions. The cases also notably include metadata on the rank and validation level of the authors of content, which are used in evaluation (wai Yim et al., 2024b). For evaluation, the competition uses a version of the deltaBLEU (Galley et al., 2015) metric to allow a single score to be computed based on word matching, weighted by the consistency (most frequent response) and the seniority of the medical professional across all responses given for that particular case. (wai Yim et al., 2024a)

The query text and target responses are given in multiple languages, namely Chinese, English, and Spanish (wai Yim et al., 2024b). It's worth noting that while the test and validation sets were translated by medical professionals, the training set of 842 cases seems to be translated automatically with some potential room for errors. For our submission we focus on only providing the English solution.

3 Related Work

There has recently been a substantial amount of interest in medical applications of multimodal machine learning, and large multimodal models. Some notable examples of research in this area include the open source LLAVA-MED model (Li et al., 2023), and ELIXR, with the latter, similar to our work, exploring not only the application of large multimodal models, but also training a model using CLIP (Xu et al., 2023). However, while there has been significant focus on certain areas such as radiology, the area of dermatology has not been explored to the same extent. Cirone et al. notably found that GPT-4V could accurately differentiate between benign lesions and melanoma (Cirone et al., 2024). However, this is a much less challenging task than the one proposed in this shared task, as the problem space is much smaller in scope than responding to dermatology questions which are not necessarily in the train set, with even the conditions of interest not necessarily being in the train set. The limited performance of our solution, along with it being by far the best performing solution in this competition demonstrate the challenge of this task, and highlight the need for significant progress before deployment in a clinical setting. However, our work highlights potentially important directions for future research, including further investigation of multi-stage LLM systems, and the importance of evaluation metrics in the benchmarking of the clinical efficacy of developed systems.

Rank	Team	dBLEU (English)
1	WangLab	12.855
2	kiyoonyoo	3.827
3	amdada	2.662
4	romarcg	2.133
5	xiaolihaixiao	1.758
6	pvashisht	0.923
7	nadia	0.717
8	abrygo	0.457

Table 1: Top 8 teams' performance on the English category for the MEDIQA-M3G competition

4 Results

Upon examination of the evaluation metric and competition data, we have determined that a short response focusing on disease diagnosis alone is the most advantageous. This is due to two reasons. First, we notice both the training and validation sets often contain short responses, and in many cases merely the skin condition presented in the associated images. Second, the evaluation metric's penalty for short responses is significantly smaller than a longer, partially correct response. Given these initial findings, we evaluated two methods as outlined in 1 which took 1st and 2nd place in the English category of the leaderboard during the MEDIQA-M3G challenge by a significant margin over the next best submitted solution, the latter of which received a deltaBLEU score of 3.827 during the competition. The methods will be elaborated in the following sections in detail.

4.1 Claude 3 Opus API solution

The higher scoring of the two methods during the competition consists of two successive API calls to Claude 3 Opus (Anthropic). For each case in the test set, the first API call generates possible differential diagnosis for the given images, and the second API call further processes the response into the name of the most likely disease only, which is then returned.

This exact configuration was decided based on trial and error. Table 2 outlines the solutions tested. Notably, we observe that the disease diagnosis given by Claude 3 Opus was poorer quality when

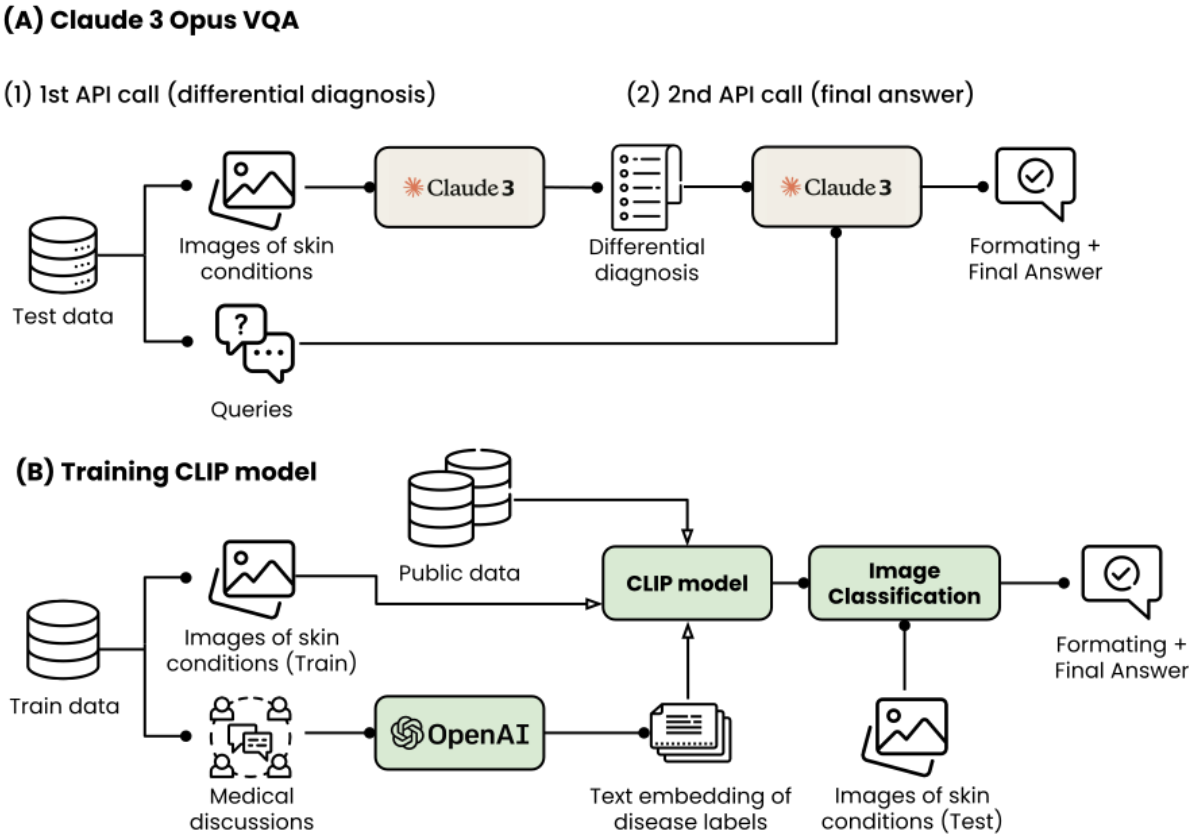

Figure 1: Overview of the two winning solutions. A) Test cases are directly submitted to the Claude 3 Opus API. The first of the two consecutive API calls generates differential diagnosis using only the images in the test cases and the second API call optionally includes the associated queries, specifies formatting, and generates final answer. The top performing Claude 3 Opus solution did not utilize test queries. B) The medical discussions included as a part of the training data is used to extract the most likely disease label for each case using GPT4-Turbo from OpenAI. The resulting image-disease label pair are used in conjunction with publicly available data to train a joint embedding in the style of CLIP. The disease labels are embedded using EmbeddingV3 from OpenAI and used to train the image encoder (ResNet50) and both the image and text projection layers. Finally, once the model is trained, the test images are classified inside the learned joint embedding which becomes the final output before performing post processing.

the prompt constrains the output format upon manual review. This was further confirmed by the inferior performance of the 1-call result. Therefore, we let the API generate differential responses with the provided images alone without any constraints on the format of the first response, and use a second API call to reformat that response into the desired form, which is just the name of the skin condition without any abbreviations.

Furthermore, we observe that including the accompanying query text for each case either in the 1st or 2nd pass was not able to outperform simply using the image alone to make the predictions. This finding may be attributed to the inconsistent information present in the query text, which may often harm the prediction from Claude 3 in some cases. It may also be a potential limitation of Claude's ability to reason with text and image simultaneously. Indeed the resulting predictions had substantial room for improvement even under the most favorable setting tested. All prompts used to produce the solutions in Table 2, including the winning solution are outlined in Appendix.

4.2 CLIP image classification solution

The second solution we've explored took second place during the MEDIQA-M3G challenge, and with subsequent tuning after the competition, was able to overtake the Claude 3 Opus API solution under the same evaluation setting used during the competition. The CLIP based solution involved learning a joint representation between the images of the skin conditions and their accompanying disease label. We achieved this by using a contrastive learning setup inspired by CLIP (Radford et al., 2021). We use a ResNet-50 (He et al., 2015) en-

Table 2: Performance of various Claude 3 Opus based solutions. 1 Call involves simply generating a response based on images, whereas 2 Calls involve first generating a differential diagnosis, then using a second API call to come up with a final diagnosis. For Img+text, both modalities are used in the first API call to generate the differential, whereas for Img then text the first API call uses only images, then the second API call uses text

Scen.	dBLEU	BP	Ratio	Hyp_len	Ref_len
Img (1 Call)	10.529	0.984	0.984	498	506
Img (2 Calls)	12.855	0.994	0.994	485	488
Img then text (2 Calls)	10.905	0.983	0.983	527	536
Img + text (2 Calls)	10.905	1.000	1.004	523	521

coder initialized with pretrained weights as the image encoder. Image augmentations include random flip, random rotations, random spatial cropping and random contrast adjustments to improve diversity and robustness of training. To obtain the disease label from the provided responses of medical professionals, we input all medical professional responses for each case in a GPT4-Turbo API call and prompt the GPT4-turbo model to return the most consistent disease diagnosis among all responses. We also curate additional image-disease pairs (n = 25528) in the domain of dermatology from publicly available sources. It's worth highlighting that there were 1245 unique disease labels among the image-disease label pairs curated. The label sparsity effectively makes training a traditional supervised classification model difficult. However, we make the observation that these labels were often the result of label inconsistency and frequently shared semantic meaning, which motivated our use of OpenAI's EmbeddingV3 (OpenAI, 2024) model to produce consistent, semantically meaningful word embeddings which effectively serve as the text encoder in our CLIP learning framework. We visualize the embeddings of the disease labels and verify that indeed many diseases with similar descriptions cluster together, as evident in Figure AS2. Specific hyperparameters used to produce the highest scoring CLIP solution are outlined in Table 6.

4.2.1 Image classification via nearest neighbour retrieval

Once the image encoder and the respective image and text projection layers are trained, the resulting joint embedding can be used to perform image classification via nearest neighbour retrieval. Specifically, we embed each image associated with a given case in the competition test set and find 5 nearest neighbours for each embeded image. We test 4 different conditions, namely retrieval between the image embedding of the query (testing dataset) and either its nearest 5 text or image embeddings from the reference (training dataset), and whether the nearest neighbours are computed in PCA space (10 components) or as normal. We then pool the labels associated with the retrieved examples via majority voting and return the final predicted label for the case. The resulting scores are presented in Table 4. Of note, during the competition (1st row), random augmentations were mistakenly not turned off during inference when obtaining the image embeddings. This did not lead to better performance and was corrected after the competition concluded.

4.2.2 Importance of batch size

The CLIP loss heavily relies on a diverse source of positive and negative pairs to converge to a good solution. It's often the case that bigger batch sizes give more robust joint representations. However, under low data settings such as for this competition where the available labelled data is scarce, larger batch sizes may lead to overfitting which is destructive for generalization. We test 3 different batch sizes ranging from 128 to 512 and observe that a batch size of 256 is most suitable under prior evaluation scripts shared by the competition organizers. However, when using the updated evaluation script during test phase of the challenge, we observe that both batch size 256 and 512 exhibit comparable performance. The results under the updated evaluation script are presented in Table 3.

Table 3: Performance of the CLIP based solution across different batch sizes

Model	dBLEU	BP	Ratio	Hyp_len	Ref_len
CLIP (batch 128)	10.434	0.980	0.980	483	493
CLIP (batch 256)	12.080	0.966	0.966	461	477
CLIP (batch 512)	12.289	0.983	0.984	447	485

Table 4: Performance of CLIP with different retrieval related strategies, including retrieval in the PCA space (n=10), and retrieving based on either the image or the text embedding of the reference. The first row indicate the CLIP based solution submitted during the competition. Of note, the random image augmentations during inference were enabled unintentionally during the competition but disabled for all subsequent experiments.

Random. Aug	PCA Space	Query-Reference	dBLEU
Yes*	Yes	Image-Image	11.979
No	No	Image-Text	12.123
No	Yes	Image-Text	8.396
No	No	Image-Image	15.884
No	Yes	Image-Image	12.079

4.3 Post processing

Post processing is performed on both the Claude 3 Opus API solution and the CLIP based image classification solution in the same way. This includes putting the output disease name in predetermined sentence format to mimic the style of the given responses from medical professionals, specifically in the form of "It is [Disease name].". While a naive approach to the VQA task, we find this simple formatting allows our disease labels produced from images alone to score quite competitively under the deltaBLEU evaluation metric provided by the competition organizers compared to simply returning the disease name itself as evident in Table 5.

Furthermore, unlike other competitors' solutions based on finetuning existing VQA models (such as LLaVA-med) simultaneously using both the images and the associated query text, our solution does not take advantage of any potentially useful information included in the query text. As a naive way of overcoming this limitation, we compiled a dictionary of disease names present in the training data and do exact word matching with the query text. Cases where the query text matches with the dictionary will have their model predictions replaced with the matched disease condition. These matches constitute 15 cases out of 100 in the testing data. While this naive heuristic often times do not produce the correct diagnosis, considering the difficulty of the task this approach does confer some improvement in overall deltaBLEU score. The ablations of the post processing is outlined in Table 5.

Solution	Word Matching	Sentence Structure	Both
Claude Solution	4.903	6.202	12.855
CLIP Solution (competition)	2.386	3.253	11.979
CLIP Solution (batch 256)	3.255	10.923	15.884

Table 5: Result of ablations on performance of top performing solutions. Sentence structure involves placing the predicted disease labels in predetermined sentence format, whereas word matching is a heuristic employed to utilize provided text via naively matching disease names with the given queries.

HyperParameter	Value
Image encoder	Resnet50
Projection dim	256
Batch size	256
Text embedding dim	3072
Image embedding dim	2048
Num. projection layers	1
Augmentations	RandFlip, RandRotate, RandSpatialCrop, RandAdjustContrast
Weight decay	0.001
Learning rate	0.001

Table 6: Hyperparameters corresponding to the highest performing CLIP solution

5 Discussion

We have presented two solutions to the MEDIQA2024-M3G competition, one involving API calls to an existing state of the art multimodal language model and the other involving the learning of an image-disease label joint embedding space for disease classification.

The superior performance of using two separate API calls to Claude 3 Opus over one pass was interesting to observe. The increase in performance is likely attributed to the reduced ability for the model to simultaneously reason with the images while adhering to the added difficulty of only returning the disease label without any additional textual generation. This finding is somewhat consistent with how chain of thought reasoning can improve model performance by asking the model to first consolidate evidence present in the given image followed by making several differential diagnoses. Further research such as (Zhang et al., 2023) also highlight the importance of using two-stage frameworks for multi-modal chain of thought that separate rationale generation and answer inference over one stage systems.

For the CLIP based solution, we find it extremely encouraging that a smaller scale model finetuned

on image-disease label pairs (n=25528) was able to outperform Claude 3 Opus (dBLEU of 15.884 vs 12.855). It perhaps demonstrates that smaller scale supervised training may sometimes outperform bigger more general purpose models for specific tasks of interest due to the advantage of training only on task specific examples. Furthermore, our additional experiments after the competition highlights the importance of proper selection of batch size and retrieval method. We observe that while CLIP effectively constructs a joint embedding space between images and their disease labels, the image embeddings and text embeddings remain as separate cluster in PCA space. As a result, we see that the nearest 5 neighbours in the text cluster for each embeded image (image-text) in the test set were much poorer in quality than those retrieved from the image cluster (image-image).

6 Limitations

While both the Claude 3 Opus API based solution and the CLIP based image classification solution achieved first and second place during the MEDIQA-M3G competition respectively, they have substantial room for improvement despite their leaderboard success.

First of all, the overall deltaBLEU score of both solutions are poor, mostly ranging from 10-15 dBLEU. The low absolute scores of the solutions really highlight the difficulty of the medical VQA task presented and the difficulty of such tasks in general. Upon examining the solutions, we observe that the models were seldom able to generate the exact name of the skin condition in question, although do a good job at identifying a disease similar in presentation or effect location (for example tinea scalp vs seborrheic dermatitis). Certainly both solutions require substantial improvements before they contribute meaningful benefits to the healthcare system in practice.

While the CLIP based solution was able to outperform our Claude 3 Opus API based solution with experiments conducted post competition, it is worth mentioning that such small scale finetuning may be less desirable as the model would have to be repurposed for new problems of interested each time. LLM based solutions have the advantage of being general purpose and do not have this issue. Furthermore, due to the tight schedules of the competition, both solutions were not explored to their full potential. We anticipate there are bigger upsides for the Claude 3 Opus API solution via more sophisticated prompting or compiling. Our rather simple implementation of the Claude based API solution may not represent the LLM's full capability but rather offers a competitive baseline for this task.

Next, both solutions while reproducible are not stable. The Claude API may be subject to randomness during generation due to the temperature parameter or the update of internal private model weights while the CLIP solutions observed inconsistencies during retrieval where the retrieved images' labels seldom agreed with each other despite relatively similar appearances, leading to low confidence in the final output. Retraining the CLIP model with the same experimental setup but initializing differently may yield completely different final disease label classification due to this inconsistency.

Lastly, the two solutions were formulated with the competition evaluation metric in mind as they are both framed as a disease label prediction task rather than a more usual VQA task which could cover a broader range of topics in their generated responses such as differential diagnoses, treatments and other recommendations as present in the actual ground truths for this competition. This is further reason to treat the performance of the presented solutions with a grain of salt. Specifically, upon our initial exploration, the deltaBLEU metric defined by the competition organizers favors short responses given the relatively heavy penalty incurred on incorrect k-mers present and relatively low penalty on a incomplete answer in comparison. This discourages model exploration during text generation and potentially penalizes model predictions that are correct semantically but are either too long or not containing the exact words present in the ground truths. This is highlighted in the ablation results in Table 5. Furthermore, the naive word matching often gave incorrect diagnosis as the patient writing the query does not have medical background, however the solution containing the disease label still scored well under the current metric as medical professionals respond with "not [disease label]" which has the opposite semantic meaning but similar k-mer composition. We recommend the organizers to slightly modify the existing metric to be more lenient with assessing the produced solutions and perhaps add a semantics component in addition to a k-mer based evaluation metric such as GPTscore (Fu et al., 2023), that can

provide more robustness in assessing the quality of generated responses.

Nevertheless, the competition serve as an important step towards the goal of automatically generating clinical responses given textual queries and associated images, and we sincerely thank the organizers for the work curating this dataset and organizing the competition.

7 Conclusion

We present two solutions to the English category of the MEDIQA2024-M3G shared task for Multilingual and Multimodal Medical Answer Generation. The Claude 3 Opus API based solution and the CLIP image classification based solution scored 1st and 2nd, respectively among all submissions. While there is still substantial room for improvement for these two solutions, we share and discuss our findings to contribute towards the important goal of automatically generating clinical responses given textual queries and associated images.

8 Acknowledgement

We extend our sincere thanks to the Digital Research Alliance of Canada for their support and computing resources. We also would like to express gratitude to both internal and external reviewers for their insightful feedback, which enhanced earlier versions of this paper. Finally, we would like to thank the organizers for all the work put into hosting this interesting and challenging competition.

References

Anthropic. The claude 3 model family: Opus, sonnet, haiku.

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.

Katrina Cirone, Mohamed Akrout, Latif Abid, and Amanda Oakley. 2024. Assessing the utility of multimodal large language models (GPT-4 vision and large language and vision assistant) in identifying melanoma across different skin tones. *JMIR Dermatol*, 7:e55508.

Greg Corrado and Yossi Matias. 2023. Multimodal medical AI. https://blog.research.google/2023/08/multimodal-medical-ai.html. Accessed: 2023-12-4.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as you desire.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

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.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications*, 15(1):654.

OpenAI. 2024. New embedding models and API updates. https://openai.com/blog/new-embedding-models-and-api-updates. Accessed: 2024-4-11.

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. 2021. Learning transferable visual models from natural language supervision.

Sachin Sharma, Raj Rawal, and Dharmesh Shah. 2023. Addressing the challenges of AI-based telemedicine: Best practices and lessons learned. *J. Educ. Health Promot.*, 12:338.

Julia Shaver. 2022. The state of telehealth before and after the COVID-19 pandemic. *Prim. Care*, 49(4):517–530.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

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*.

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, Jungyeon Park, Patricia Strachan, Yun Liu, Chuck Lau, Preeti Singh, Christina Chen, Mozziyar Etemadi, Sreenivasa Raju Kalidindi, Yossi Matias, Katherine Chou, Greg S Corrado, Shravya Shetty, Daniel Tse, Shruthi Prabhakara, Daniel Golden, Rory Pilgrim, Krish Eswaran, and Andrew Sellergren. 2023. ELIXR: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal Chain-of-Thought reasoning in language models.

A Claude 3 Opus API prompts

Example prompts used to perform API calling in the Claude 3 Opus solution and other tested variants.

A.1 Image only 1-call

System: You are an expert assistant to a blind dermatology student, help him identify exactly what conditions would be included in the differential for this condition? Be concise. After brief description of the images and explanation of your choice, give the most commonly occuring skin disease out of the differentials at the end and nothing else, in the form of

Answer: [Disease Name]

Content: IMG_ENC00908_00001.jpg, IMG_ENC00908_00002.jpg
Output: Answer: Dyshidrotic eczema

A.2 Image only 2-calls

System: You are an expert assistant to a dermatology student, help him identify exactly what skin conditions would be included in the differential for the images presented. Consider both resemblence and prevalence.
Content: IMG_ENC00908_00001.jpg, IMG_ENC00908_00002.jpg
Output1: Based on the images provided, the key skin findings are ... The differential diagnosis for these lesions would include:

1. Hand eczema (dyshidrotic eczema) ...
System: You are an expert assistant to a dermatology student. Given the following differentials, only return the name of the most likely diagnosis and nothing else. Do not include alternative names of the differential in brackets.
Content: Based on the images provided, the key skin findings are ... The differential diagnosis for these lesions would include:

1. Hand eczema (dyshidrotic eczema) ...
Output2: hand eczema

A.3 Image then text 2-calls

Of note, the first API remains the same to the Image only 2-calls case, but the added Additional Information field contains the text query associated with each case in the test set.

System: You are an expert assistant to a dermatology student, help him identify exactly what skin conditions would be included in the differential for the images presented. Consider both resemblence and prevalence.
Content: IMG_ENC00908_00001.jpg, IMG_ENC00908_00002.jpg
Output1: Based on the images provided, the key skin findings are ... The differential diagnosis for these lesions would include:

1. Hand eczema (dyshidrotic eczema) ...
System: You are an expert assistant to a dermatology student, given the following differentials discussed and some additional information provided, only return the name of the most likely diagnosis and nothing else. Do not include alternative names of the differential in brackets.
textbfContent: Differentials:

Based on the images provided, the key skin findings are ... The differential diagnosis for these lesions would include:

1. Hand eczema (dyshidrotic eczema) ...
Additional information: Picture 1: On the outside of the thigh, there is a small circle of lump. Approximately 2 months.
Picture 2: Small red spots on the palm. There is slight numbness in the palm. **Output2:** hand eczema (dyshidrotic eczema)

A.4 Image + text 2-calls

System: You are an expert assistant to a dermatology student, help him identify what skin conditions would be included in the differential for the presented images and additional information provided by the medical professional. If any skin conditions are mentioned in the additional information, include them as the most likely differential.
Content: Additional information: Picture 1: On the outside of the thigh, there is a small circle of lump. Approximately 2 months.
Picture 2: Small red spots on the palm. There is slight numbness in the palm.
IMG_ENC00908_00001.jpg, IMG_ENC00908_00002.jpg
Output1: Based on the provided images and additional information, here are the potential skin conditions to consider in the differential

diagnosis: ... **System:** You are an expert assistant to a dermatology student. Given the following differentials, only return the name of the most likely diagnosis and nothing else. Do not include alternative names of the differential in brackets.
textbfContent: Based on the provided images and additional information, here are the potential skin conditions to consider in the differential diagnosis:
...
Output2: picture 1: lipoma. picture 2: palmar erythema

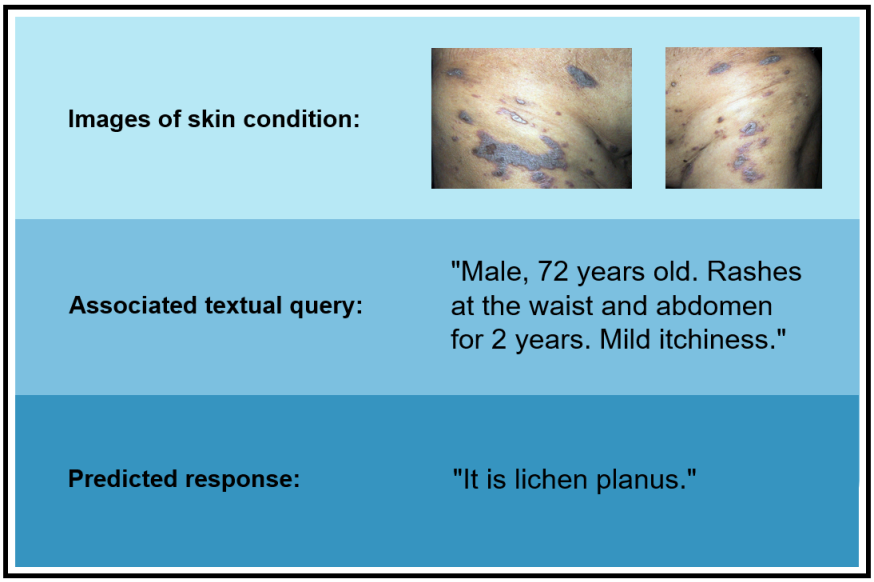

Figure S1: Representative case example illustrating the images of the skin condition, their associated textual query and the predicted response given.

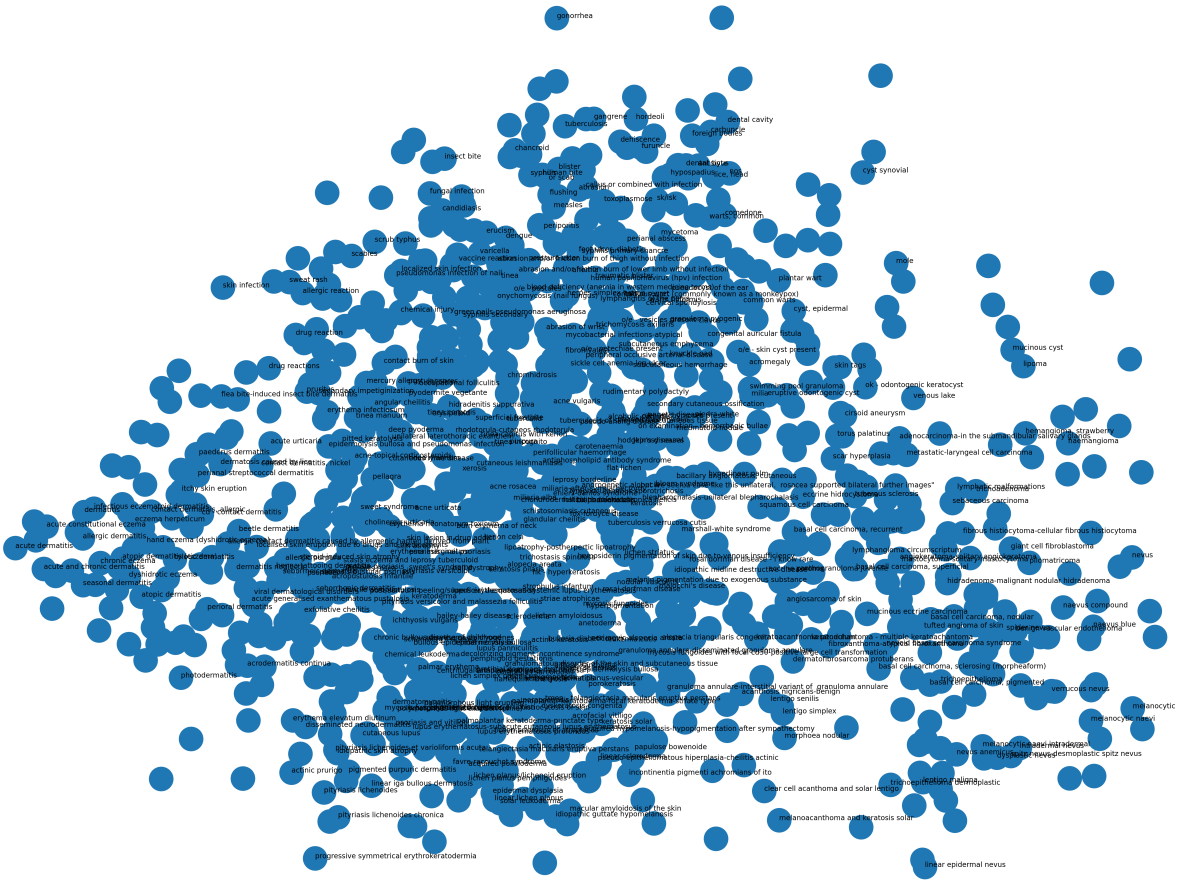

Figure S2: PCA visualization of all the training disease labels embedded by the EmbeddingV3 model. Skin conditions that are semantically similar are clustered together in this representation space.