Overview of the MEDIQA-MAGIC Task at ImageCLEF 2024: Multimodal And Generative TelemedICine in Dermatology*

Notebook for the ImageCLEF Lab at CLEF 2024

Wen-wai Yim[1], Asma Ben Abacha[1], Yujuan Fu[2], Zhaoyi Sun[2], Meliha Yetisgen[2] and Fei Xia[2]

[1]*Microsoft Health AI, Redmond, 98052, USA*
[2]*University of Washington, Seattle, 98109 USA*

Abstract

Multimodal processing and language generation require models to internally represent both language and vision, and then generate contextually appropriate responses. To do so with arbitrary images and textual inputs in the medical field, requires additional high performance and fidelity. This paper presents the overview of the MEDIQA-MAGIC shared task at ImageCLEF 2024. In this dermatological visual question-answering (VQA) task, participants receive the input of an image and a textual consumer health query, and are expected to output a textual medical answer. A total of twenty two runs were submitted with a variety of general language-vision models and fine-tuned models, with the best team achieving 8.969 BLEU points. We hope that the findings and insights explored here will inspire future research directions to support improved patient care.

Keywords

Visual Question Answering, Response Generation, Dermatology

1. Introduction

Partially in effect after the adoption of meaningful use requirements in the United States, the ability to message doctors and receive care remotely on patient portals have skyrocketed since the 2019 COVID pandemic [1, 2]. Furthermore, the establishment of online tele-health companies outside of traditional hospital settings, e.g. teledoc [3], icliniq [4] and amazon clinic [5], harkens consumer health needs for on-demand medical care access. Asynchronous online dermatology consultation is one application area for this new care delivery method. In this scenario, patients may provide their dermatology images and questions through electronic messaging; medical doctors may then likewise provide electronic responses related to treatment and medication. However, whether as an extended branch of a hospital institution or as a standalone online care alternative, these services require a medical doctor in the loop to deliver safe, reliable care – putting additional demands on provider workload.

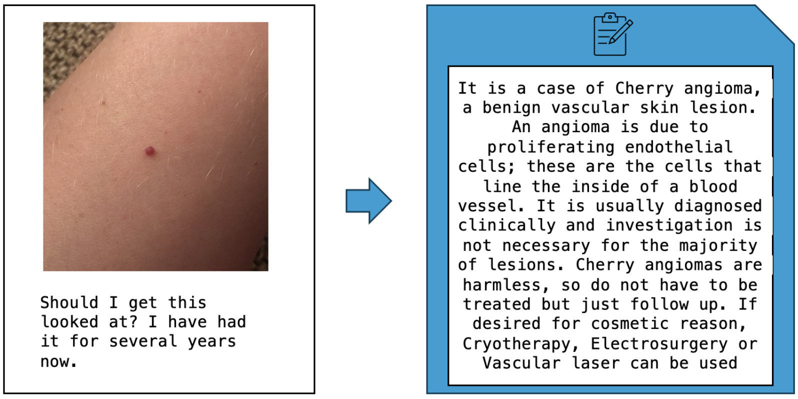
Automated models have the opportunity to provide response suggestions, which in turn may help optimize care quality and efficiency, deburdening healthcare workers. While large multi-modal language models have made significant strides, such as with the results of Gemini [6] and GPT-4o [7], there remain questions about the applicability of such models to real-world unconstrained tasks. The area of multi-modal consumer health question-answering is a challenging task. It requires processing of uncontrolled user-generated images, featuring variable angles, lighting, and resolution, as well as arbitrary textual content and queries. As a medical task, reasonable accuracy and reliability are critical.

To benchmark the state-of-the-art large multi-modal language models' performance for this problem, we have conducted the MEDIQA-MAGIC shared task as part of ImageCLEF 2024 [8]. Specifically, participants are required to compose automatic responses to patient dermatological queries with a textual question and image context. Previous related shared tasks have featured radiology-related visual question answering (VQA) [9, 10], as well as text-only consumer health answer generation [11]. Other medical VQA problems include images from the areas of pathology and GI-tract [12, 13]. Meanwhile, single modality dermatological image classification tasks include automatic categorizations

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

to predefined diseases or symptom characteristics [14, 15]. This year's edition tackles answer generation for multi-modal consumer health question task. A similar task was part of the related NAACL 2024 ClinicalNLP challenge MEDIQA-M3G 2024 [16] featuring a different dataset.

In the following sections, we introduce the tasks, describe the evaluation, present the participating teams' results, as well as provide some insight into future directions.



Should I get this looked at? I have had it for several years now.

It is a case of Cherry angioma, a benign vascular skin lesion. An angioma is due to proliferating endothelial cells; these are the cells that line the inside of a blood vessel. It is usually diagnosed clinically and investigation is not necessary for the majority of lesions. Cherry angiomas are harmless, so do not have to be treated but just follow up. If desired for cosmetic reason, Cryotherapy, Electrosurgery or Vascular laser can be used

Figure 1: In this task, a consumer health query and an accompanying image are given. The expected output is a suitable response from a dermatology medical doctor.

2. Task Description and Dataset

In this shared-task, an input instance is composed of a single textual query and an image, representing a consumer health question. The expected output is a free-text response, representing a possible doctor's answer to the query. Figure 1 shows an example instance.

The dataset was sourced from real consumer health queries found on Reddit for posts related to dermatological problems (subreddit r/DermatologyQuestions) [1]. Encounters were filtered out if they met at least one of the following exclusion criteria: (a) images that included identifying features (e.g. full faces), (b) queries that were not seeking information (e.g. "look at my tatoo"), (c) images including genitalia, and (d) images that contained annotations (e.g. drawn arrows). Gold standard responses were generated by 3 certified practicing dermatologists. The train and validation sets were single annotated; the test set was double-annotated. To comply with Reddit data usage guidelines, only post IDs and our response labels were shared with participants. Participants who registered through Reddit could receive API credentials to access Reddit's data. Afterwards, the participants could use the supplied download script[2] to retrieve the original input data.

Table 1 provides summary statistics of the dataset. A single query may involve multiple anatomic locations. Because users may delete content, the final set of test set IDs was determined by the subset of test IDs retrieval shortly after the submission deadline. As a consequence, the final number of test set encounters may include fewer instances than the original labeled test set. The data here used a subset of the DermaVQA dataset, for which the full corpus creation description can be found in [17].

3. Evaluation Methodology

We evaluated the system responses by comparing them with the double-annotated gold standard responses per query. We used relevant multi-reference metrics/variants including:

[1] https://www.reddit.com/r/DermatologyQuestions/
[2] https://github.com/wyim/MEDIQA-MAGIC-2024

Table 1
DermaVQA Reddit Subset Data Characteristics. During the challenge, the final number of test set encounter may drop due to content deletions. The final encounters for shared task test evaluation (test-final) were determined by available retrievable posts shortly after submission closed. Thus, test-final is a subset of test. Total includes the statistics of the aggregated train, valid, and test sets.

	train	valid	test	test-final	total
FREQUENCIES					
encounters	347	50	93	78	490
encounters with 1 response	347	50	0	0	397
encounters with 2 response	0	0	93	78	93
LENGTH					
mean query len	30.8	28.1	28.6	29.4	30.1
mean answer len	93.6	94.7	96.3	97.0	94.6
LOCATIONS					
arm	44	7	13	11	64
back	25	5	5	4	35
chest	31	3	6	4	40
foot	32	1	11	11	44
hand	51	11	12	11	74
head	126	16	33	30	175
leg	33	6	14	10	53
unk	28	3	9	7	40

deltaBLEU. deltaBLEU is a variant of SacreBLEU developed for response generation, in which many diverse gold standard responses are possible [18]. The metric incorporates human-annotated quality rating and assigns higher weights to n-grams from responses rated to be of higher quality. The authors have shown this method produces higher correlation with human rankings compared to previous BLEU metrics. In this task, we weigh both annotator responses as equal, defaulting to a normal BLEU score behavior. This metric was used for the shared task ranking.

BERTScore. BERTScore[3] [19] averages the maximum word embedding similarity scores between two texts based on BERT embeddings. This metric has been shown to work well on a variety of tasks, including image captioning and machine translation. The maximum was taken over multiple over pairwise scores when multiple references were available.

MEDCON. In this task, we propose MEDCON a medical information-extraction-based metric. The metric uses QuickUMLS[4] to identify medical concepts in conjunction with an in-house Llama-based assertion classifier [20]. Concepts identified by QuickUMLS are normalized according to a curated concept map. Precision, recall, and F1 were calculated based on combined concept and assertion statuses. The maximum was taken over multiple pairwise scores when multiple references were available. A variant of this metric, excluding the assertion status, was used in our previous work for measuring clinical note summarization from medical dialogue[21]. The evaluation code can be found in our GitHub repo[5].

4. Results

Out of 30 initial registrations, 3 teams submitted results in a total of 22 runs. The final results are shown in Table 3. The submitted systems represented a variety of solutions, including leveraging out-of-the-box Gemini [6] models (YuanAI), applying small visual language models (VisionQAries), and utilizing visual-language encoders with cosine similarities (IRLab@IIT_BHU). The ranges of scores performed at the lower spectrum for all three metrics (100 total for BLEU, and 1.0 for BERTScore and

[3] github.com/Tiiiger/bert_score
[4] github.com/Georgetown-IR-Lab/QuickUMLS
[5] https://github.com/wyim/MEDIQA-M3G-2024

Table 2
MEDIQA-MAGIC 2024: Participating teams, number of runs, submitted codes, and working notes papers.

team	affiliation	runs	paper
IRLab@IIT_BHU	IIT (BHU), Varanasi, India	11	[22]
VisionQAries	Poland	8	[23]
YuanAI	Yuan Ze University, Taiwan	3	[24]

Table 3
Performance of the participating teams in the MEDIQA-MAGIC 2024 Answer Generation Task. Rank is based on the BLEU score.

team	MODELS_EXACT	BLEU	BERTScore	MEDCON	rank
VisionQAries	moondream2	8.969	0.844	0.077	1
IRLab@IIT_BHU	Clip, Cosine similarity	4.536	0.839	0.066	2
IRLab@IIT_BHU	CLIP, Cosine similarity, data augmentation	4.490	0.840	0.055	3
YuanAI	llama3,gemini-pro	4.371	0.856	0.087	4
IRLab@IIT_BHU	Clip, triplet loss, textgenie	4.155	0.839	0.06	5
IRLab@IIT_BHU	CLIP, Cosine similarity, textgenie	3.986	0.838	0.052	6
IRLab@IIT_BHU	CLIP, BIsltm	3.951	0.839	0.075	7
YuanAI	llama	3.939	0.842	0.085	8
YuanAI	llama3,gemini-pro	3.939	0.842	0.085	9
VisionQAries	moondream2	3.310	0.841	0.106	10
VisionQAries	moondream2	2.749	0.837	0.111	11
IRLab@IIT_BHU	CLIP, GPT2-xl	2.447	0.831	0.066	12
IRLab@IIT_BHU	BERT, Clip, BIlstm	2.267	0.838	0.046	13
IRLab@IIT_BHU	CLIP, GPT2-xl	2.215	0.825	0.059	14
VisionQAries	moondream2	2.211	0.829	0.088	15
IRLab@IIT_BHU	CLIP, GPT2-xl	2.140	0.841	0.073	16
VisionQAries	tiny-llava-v1-hf	1.945	0.842	0.106	17
VisionQAries	moondream2	1.570	0.835	0.099	18
VisionQAries	tiny-llava-v1-hf	1.472	0.837	0.087	19
VisionQAries	moondream2	1.250	0.839	0.100	20
IRLab@IIT_BHU	CLIP, GPT-2	1.008	0.840	0.052	21
IRLab@IIT_BHU	Clip, Cosine similarity	0.400	0.831	0.023	22

MEDCON), indicating the difficulty of the task. Each teams' system descriptions are described in the following paragraphs:

IRLab@IIT_BHU This team's approach involved multiple steps, using both pre-trained language and vision-language models in conjunction with their own neural network architecture. The team first manually labeled instances into 160 hand-crafted, non-mutually-exclusive categories to be used as targets. For each query instance, image and text were passed through a CLIP [25] vision and text encoder respectively. Text-encoded data was then sent through a Bi-LSTM, and the vision-encoded data, passed through an MLP layer. The results of both were averaged to produce a label vector. During training, the vector is compared with positive and negative label embeddings using a weighted cosine similarity loss. During inference, the combined embedding was compared with the closest label embedding and assigned the corresponding class. The team also experimented with data augmentation by adding paraphrased versions of the original data using TextGenie [26], and using GPT2 [27] as the final classifier.
VisionQAries This team's solutions focused on applying small multi-modal models. Mainly, they tested two different approaches: (a) direct prompting on pre-trained moondream2 [28] and TinyLLaVA [29] models; and (b) fine-tuning moondream2 model. During testing, they compared two different prompts. Based on their results, they found that fine-tuning models gave better results than direct prompting in the context of BLEU scores.

YuanAI This team used a two-step approach. In the first step, they utilized the Gemini as their image-2-text model to create a descriptive text. In a second step, using the textual query and input from the previous step, they employed a LoRA fine-tuned Llama3 [30] to use the output from the previous step and the query as input to an LLM model to generate the final response.

As evidenced in the spread of scores of each team across the rankings, the same setups with small variations may produce widely different performances. For example, Team VisionQAries' moondream2 runs had rankings at 1, 10, and 20 at BLEU scores of 8.969, 3.310, and 1.250 – utilizing different fine-tuning and prompting variations. Although the best system scored by BLEU was much higher compared to the next best system, the other metrics did not reflect this difference.

5. Discussion and Conclusions

Multi-modal question answering for unconstrained answers is a challenging problem. From the similarity of scores in a wide variation of systems, it is clear that no one architecture is particularly superior at this task.

This year's related 2024 NAACL ClinicalNLP MEDIQA-M3G [16] task posed the same problem of dermatological multi-modal answer generation as this shared task, however the data characteristics of these two challenges were completely different. The NAACL ClinicalNLP MEDIQA-M3G shared task utilized the iiyi subset of the DermaVQA dataset [17]. Particularly, the data was sourced form a Chinese medical platform where most query posts contained 1-3+ images and the naturally-occurring responses tended to have shorter replies. In contrast, in this challenge, the MEDIQA-MAGIC shared task, data originated from Reddit posts, with one image and brief shorter queries; where responses were generated by medical doctors hired for the dataset creation - often providing very complete well-formed answers. As a consequence, in this task, the data contained shorter queries and longer answers with an average of 30 and 95 words, respectively, compared to 80 and 12 words in the MEDIQA-M3G shared task English version data subset. Although the magnitude of BLEU and BERTscores were similar in both challenges, the MEDCON scores were much lower (highest 0.1 F1 in this task, compared to a high of 0.29 in the M3G task). This can be attributed to the long answers expected in this dataset which may include many more concepts. That said, the overall modest scores across both shared tasks highlight the need for improved answer generation methods for dermatological VQA. More details on the dataset can be found in our dataset paper[17] and released dataset(https://osf.io/72rp3/).

Despite differences, similar task-related issues arose in both of these shared tasks. Firstly, true dermatology gold standards for benign maladies are rare as typically suspected malignant lesions are prioritized for pathologically testing. This may lead to differences in dermatological expert opinions which cannot be resolved. Even with access to large private health records, methods to tackle the determination of the best gold label will require additional investigation. Secondly, in contrast to previous VQA tasks, this task expected long-form natural language outputs. Although prior VQA datasets had textual outputs, in reality, the number of question types are limited, with answers on average at 1-2 words long. In fact, all previous VQA tasks report accuracy as a metric. Natural language generation evaluation with respect to VQA is an area needing much more future research, particularly with respect to fairly evaluating instances with multiple diverse possible answers and evaluating instances of long free-text responses.

This shared task revealed a multitude of opportunities for future modeling, corpus creation, and evaluation. Future directions of study includes testing additional vision-language models, incorporating intermediate image segmentation or image extraction steps, and re-ranking answers. In the future, these can be tested on the larger combined DermaVQA dataset[17]. In future editions of this task, we will experiment with other evaluation methods, e.g. ranking or weighting based on normalized medical concepts. We hope that the benchmarks, insights, and datasets presented here will inspire future research directions.

Acknowledgments

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

References

[1] T. F. Bishop, M. J. Press, J. L. Mendelsohn, L. P. Casalino, Electronic communication improves access, but barriers to its widespread adoption remain, Health affairs (Project Hope) 32 (2024) 10.1377/hlthaff.2012.1151.

[2] C. A. Sinsky, T. D. Shanafelt, J. A. Ripp, The electronic health record inbox: Recommendations for relief 37 (2024) 4002–4003.

[3] teledoc, https://www.teladochealth.com/, 2024. Accessed: 2024-05-31.

[4] icliniq, https://www.icliniq.com/, 2024. Accessed: 2024-05-31.

[5] Amazon clinic, https://clinic.amazon.com/, 2024. Accessed: 2024-05-31.

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

[7] Gpt-4o, https://openai.com/index/hello-gpt-4o/, 2024. Accessed: 2024-05-31.

[8] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[9] J. J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, Scientific data 5 (2018) 1–10.

[10] A. Ben Abacha, S. A. Hasan, V. V. Datla, D. Demner-Fushman, H. Müller, Vqa-med: Overview of the medical visual question answering task at imageclef 2019, in: Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes, 9-12 September 2019, 2019.

[11] A. Ben Abacha, E. Agichtein, Y. Pinter, D. Demner-Fushman, Overview of the medical question answering task at trec 2017 liveqa, in: TREC 2017, 2017.

[12] X. He, Z. Cai, W. Wei, Y. Zhang, L. Mou, E. Xing, P. Xie, Towards visual question answering on pathology images, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 708–718.

[13] S. Hicks, A. M. Storås, P. Halvorsen, T. de Lange, M. Riegler, V. L. Thambawita, Overview of imageclefmedical 2023 - medical visual question answering for gastrointestinal tract, in: Conference and Labs of the Evaluation Forum, 2023.

[14] R. Daneshjou, K. Vodrahalli, W. Liang, R. A. Novoa, M. Jenkins, V. Rotemberg, J. M. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong, R. Sahasrabudhe, J. Zou, A. S. Chiou, Disparities in dermatology ai performance on a diverse, curated clinical image set, Science Advances 8 (2021).

[15] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, O. Badri, Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1820–1828.

[16] A. Ben Abacha, W. Yim, Y. Fu, Z. Sun, F. Xia, M. Yetisgen, M. Krallinger, Overview of the mediqa-m3g 2024 shared tasks on multilingual multimodal medical answer generation, in: NAACL-ClinicalNLP 2024, 2024.

[17] W. Yim, Y. Fu, Z. Sun, A. Ben Abacha, M. Yetisgen, F. Xia, Dermavqa: A multilingual visual question answering dataset for dermatology, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, 2024.

[18] M. Galley, C. Brockett, A. Sordoni, Y. Ji, M. Auli, C. Quirk, M. Mitchell, J. Gao, B. Dolan, 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), Association for Computational Linguistics, Beijing, China, 2015, pp. 445–450.

[19] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, ArXiv abs/1904.09675 (2019).

[20] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).

[21] W. Yim, Y. Fu, A. B. Abacha, N. Snider, T. Lin, M. Yetisgen, Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation, Scientific Data 10 (2023). URL: https://api.semanticscholar.org/CorpusID:259075199.

[22] A. Agrawal, S. Pal, Irlab@iit_bhu at mediqa-magic 2024: Medical question answering using classification model, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[23] P. Cieplicka, J. Kłos, M. Morawski, Visionqaries at mediqa-magic 2024: Small vision language models for dermatological diagnosis, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[24] H. Fu, H. Huang, Yuanai at mediqa-magic 2024: Improving medical vqa performance through parameter-efficient fine-tuning, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021. URL: https://api.semanticscholar.org/CorpusID:231591445.

[26] Text genie, https://github.com/hetpandya/textgenie, 2024. Accessed: 2024-05-31.

[27] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[28] moondream.ai, https://moondream.ai/, 2024. Accessed: 2024-05-31.

[29] B. Zhou, Y. Hu, X. Weng, J. Jia, J. Luo, X. Liu, J. Wu, L. Huang, Tinyllava: A framework of small-scale large multimodal models, 2024. arXiv:2402.14289.

[30] llama3 model, https://llama.meta.com/llama3/, 2024. Accessed: 2024-05-31.