Overview of the MEDIQA-WV 2025 Shared Task on Woundcare Visual Question Answering

Wen-wai Yim
Microsoft Health AI
yimwenwai@microsoft.com

Asma Ben Abacha
Microsoft Health AI
abenabacha@microsoft.com

Meliha Yetisgen
University of Washington
melihay@uw.edu

Fei Xia
University of Washington
fxia@uw.edu

Abstract

Electronic messaging through patient portals facilitates remote care, connecting patients with doctors through asynchronous communication. While convenient, this new modality places an additional burden on physicians, requiring them to provide remote care as well as to see patients in clinic. Technology that can automatically draft responses for physician review is a promising way to improve clinical efficiency. Here, building on the 2024 MEDIQA Multilingual Multi-modal Medical Answer Generation (MEDIQA-M3G) challenge on dermatology, we present the 2025 MEDIQA Woundcare Visual Question Answering (MEDIQA-WV) shared task focusing on generating clinical responses to patient text and image queries. Three teams participated and submitted a total of fourteen systems. In this paper, we describe the task, datasets, as well as the participating systems and their results. We hope that this work can inspire future research on wound care visual question answering.

1 Introduction

Patient health portals enable asynchronous care, allowing patients to electronically submit clinical questions at any time. While this technology provides flexibility and convenience for both patients and clinicians, it also generates an unprecedented volume of additional work for care providers (Bishop et al.; Sinsky et al.).

In parallel, large multimodal general AI models have achieved state-of-the-art performance on benchmark tasks involving both classification and generation. However, despite these advances, such models often struggle with complex tasks requiring logical reasoning and multi-step inference. Medicine, in particular, demands integration of extensive medical knowledge, individual patient history, and current treatment protocols. Moreover, errors in this domain can have serious health consequences.

In the 2024 MEDIQA-M3G Challenge (Yim et al., 2024), we investigated the automatic generation of medical responses to patient queries in dermatology, incorporating both text and image components. In this new 2025 MEDIQA-WV challenge[1], we extend our exploration to the domain of wound care.

2 Task & Dataset Description

Participants are given a patient query that includes a textual description and one or more associated images. The objective is to create free-text responses as if they were written by a clinician.

The training set includes a single gold response, the validation set includes two gold responses, and the test set includes three gold responses.

Both English and simplified Chinese versions are available for each instance. Table 1 provides an example of an instance with multiple gold responses. Table 2 and 3 provide summary characteristics of the dataset.

The patient queries were sourced from two open Chinese community forums (tieba.baidu.com and zhidao.baidu.com). The questions were translated into English by bilingual registered nurses and a medical translator. Each query was reviewed and answered by at least one of three U.S. medical domain experts: an emergency medicine physician, an emergency medicine resident, and a general surgeon. The created gold answers were then translated back into Chinese. In addition, metadata categories such as wound type and wound thickness were single-annotated and included. A detailed description of the dataset construction can be found in this work (Yim et al., 2025).

3 Evaluation Metrics & Baselines

Evaluation used three classical NLG metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin,

[1] https://sites.google.com/view/mediqa-2025

Proceedings of the 7th Clinical Natural Language Processing Workshop, pages 17–21

PATIENT'S QUESTION & IMAGE
被猫隔着厚牛仔裤抓了一下，留下一道血痕，伤口如图， 请问需要接种狂犬疫苗吗 I was scratched by a cat through thick jeans, leaving a blood mark as shown in the picture. Do I need to get a rabies vaccine? 
GOLD RESPONSES
Annotator 1: 在这种情况下不需要注射狂犬疫苗。 No need for rabies vaccine in this situation.
Annotator 2: 如果猫的疫苗接种状态未知且无法隔离，虽然风险较低，具体取决于地区，建议接种疫苗。否则，局部伤口护理，使用局部抗生素和敷料是就行了。 If vaccination status of cat unknown and cannot be quarantined, while risk is low, depending on the area, would recommend vaccination. Otherwise, local wound care with topical antibiotic and dressing is sufficient
Annotator 3: 你应该不需要为猫抓伤接种狂犬症疫苗，因为从抓伤传播狂犬症的风险非常低。当然，如果你居住在一个狂犬症发病率很高的也区，我会另作考虑。虽然如此，我会继续观察伤口部位；如果伤口已受感染，你可能需要抗生素来治疗。我会用抗生素软膏包扎伤口。 You should not need to get a rabies vaccine for a cat scratch as the risk of transmission of rabies from a scratch is very low however, if you live in an area with high instance, I would consider it. However, I would continue to keep an eye on the area as you may need antibiotics for the cat scratch if it becomes infected. I would dress the wound with antibiotic ointment.

Table 1: Example where gold responses share similar content but differ in emphasis, level of explanation, and final recommendations (no vaccination versus vaccination), highlighting the importance of using multiple reference answers in evaluation to capture diverse opinions and perspectives.

Split	#Instances	#Responses	#Images
Training	279	279	449
Validation	105	210	147
Test	93	279	152

Table 2: Data Statistics

Split	EN		ZH	
	Query	Response	Query	Response
Training	46	29	52	43
Validation	44	41	50	60
Test	52	47	60	68

Table 3: Response Length Statistics (Mean Token Count). English tokens are per word, Chinese per character.

2004), BERTScore[2] (Zhang et al., 2020), and three LLM-as-judge variants (DeepSeek-V3, Gemini, GPT-4o). The exact configurations for the models and tokenizers are given in Table 4. For LLM-as-judge models, we used a consistent prompt for the same language. Both English and Chinese prompts for the LLM-as-judge methods are given in Table 5. Finally, we calculate the average score across all metrics.

For reference, we provide three baselines based on vision-language models with English and Chi-

[2]The mean is taken over all gold responses per instance

Configurations
BERTSCORE
github.com/Tiiiger/bert_score
tokenizer: "en"/"zh"
model: "microsoft/deberta-xlarge-mnli" for English
"zh" for Chinese
BLEU
github.com/mjpost/sacrebleu
use_effective_order: true
EN tokenizer: tokenize_13a
ZH tokenizer: tokenize_zh
DEEPSEEK
ai.azure.com
AZURE AI Foundry model ID:
DeepSeek-V3-0324
content-filter: None
Gemini
https://cloud.google.com/ai/generative-ai?hl=en
Google GenAI Model Name:
gemini-1.5-pro-002
hate-speech: Block-None
harrassment: Block-None
GPT-4o
oai.azure.com
AZURE AI Foundry OpenAI model name:
gpt-4o
content-filter: None
ROUGE
huggingface.co/spaces/evaluate-metric/rouge
tokenizer: same as BLEU

Table 4: Evaluation Metric Configurations. Defaults are used if not otherwise mentioned.

SYSTEM: You are a helpful medical assistant. USER: Given a patient {QUERY}, and a list of {REFERENCE RESPONSES}, please evaluate a {CANDIDATE RESPONSE} using a three-step rating described below. Rating: 0 - {CANDIDATE RESPONSE} is incomplete and may contain medically incorrect advice. Rating: 0.5 - {CANDIDATE RESPONSE} is incomplete but has partially correct medical advice. Rating: 1.0 - {CANDIDATE RESPONSE} is complete and has medically correct advice. The {REFERENCE RESPONSES} represent answers given by domain experts and can be used as references for evaluation. QUERY: REFERENCE RESPONSES: CANDIDATE RESPONSE: RATING:
SYSTEM: 你是一个很有帮助的医疗助手。 USER: 给定病人提出的{问题}以及一系列{参考回复}，请使用下述的 3 级评分制度来评估{待测回复}。 评分：0 - {待测回复} 不完整且与所有{参考回复}事实不符。 评分：0.5 - {待测回复} 不完整但与至少一个{参考回复}事实相符。 评分：1.0 - {待测回复} 完整且与至少一个{参考回复}事实相符 问题: {} 参考回复:{} 待测回复:{} 评分:

Table 5: LLM-as-Judge Prompts for General AI Models (English TOP, Chinese BOTTOM)

nese proficiency: Baseline 1: Gemini-1.5-pro-002, Baseline 2: GPT-4o, Baseline 3: Qwen-VL[3].

The prompts for generating the English and Chinese baseline answers were:

- English: *"Please answer as a professional medical doctor, answer limited to 41 words*[4]*. {QUERY_TITLE}: {QUERY_CONTENT}"*
- Chinese: *"请以专业医生的身份提供建议，答案只限60字 {QUERY_TITLE}: {QUERY_CONTENT}"*

4 Official Results

Three teams participated in this shared task, with a total of fourteen submissions[5]. The teams were EXL Health AI Lab (India), DermaVQA (United Kingdom), and MasonNLP (United States). Tables 6 and 7 present the official results on the English and Chinese datasets, respectively. The overall baseline averages were higher for the Chinese subset than for the English subset. Notably, only three of the fourteen submissions included systems capable of handling the Chinese subset. Two teams submitted working notes, and we provide a description of their systems below.

MasonNLP (Karim and Özlem Uzuner, 2025). This team experimented with zero-shot, few-shot, and retrieval-augmented few-shot learning approaches to identify the most relevant training data. The backbone model used was meta-llama/Llama-4-Scout-17B-16E-Instruct. The retrieval-augmented generation (RAG) component was built using semantic text embeddings from sentence-transformers/all-MiniLM-L6-v2 and vision–language embeddings from CLIP (openai/clip-vit-base-patch32). The two most similar training examples were retrieved. In experimentation, the RAG system with both image and text search achieved the highest performance.

EXL Health AI Lab (Durgapraveen et al., 2025). This team experimented with two approaches: (1) a two-step generation approach, which first classified relevant metadata (e.g., wound

[3] https://huggingface.co/Qwen/Qwen-VL

[4] Length suggestions used the average response lengths in the validation sets.

[5] A single submission may include responses in both English and Chinese.

Team	BLEU	ROUGE-L	BERTScore	DeepSeek-V3	Gemini	GPT-4o	AVG
EXL Health AI Lab	0.099	0.456	0.622	0.682	0.645	0.715	0.473
EXL Health AI Lab	0.130	0.452	0.619	0.635	0.591	0.629	0.457
EXL Health AI Lab	0.130	0.452	0.619	0.625	0.586	0.618	0.455
EXL Health AI Lab	0.057	0.456	0.623	0.607	0.629	0.667	0.45
EXL Health AI Lab	0.057	0.455	0.623	0.591	0.634	0.624	0.44
EXL Health AI Lab	0.037	0.441	0.611	0.604	0.570	0.618	0.427
MasonNLP	0.089	0.422	0.59	0.535	0.554	0.554	0.414
MasonNLP	0.073	0.433	0.604	0.589	0.565	0.532	0.411
EXL Health AI Lab	0.064	0.448	0.621	0.512	0.500	0.505	0.410
EXL Health AI Lab	0.064	0.448	0.621	0.499	0.505	0.505	0.410
DermaVQA	0.076	0.455	0.606	0.427	0.457	0.371	0.377
MasonNLP	0.047	0.235	0.325	0.321	0.301	0.339	0.236
MasonNLP	0.017	0.140	0.192	0.210	0.188	0.215	0.141
Baseline 1: Gemini-1.5-pro-002	0.064	0.449	0.621	0.791	0.817	0.683	0.571
Baseline 2: GPT-4o	0.062	0.450	0.623	0.756	0.731	0.688	0.552
Baseline 3: Qwen-VL	0.051	0.428	0.599	0.513	0.478	0.473	0.424

Table 6: Results - English. Evaluation metrics DeepSeek-V3, Gemini, and GPT-4o are reported under an LLM-as-judge setup.

Team	BLEU	ROUGE-L	BERTScore	DeepSeek-V3	Gemini	GPT-4o	AVG
DermaVQA	0.102	0.489	0.656	0.570	0.548	0.511	0.439
MasonNLP	0.000	0.006	0.008	0.011	0.011	0.011	0.006
MasonNLP	0.000	0.005	0.007	0.011	0.011	0.011	0.006
Baseline 1: Gemini-1.5-pro-002	0.118	0.501	0.661	0.941	0.957	0.898	0.679
Baseline 2: GPT-4o	0.123	0.496	0.666	0.844	0.860	0.753	0.624
Baseline 3: Qwen-VL	0.094	0.484	0.658	0.763	0.694	0.699	0.565

Table 7: Results - Chinese. Evaluation metrics DeepSeek-V3, Gemini, and GPT-4o are reported under an LLM-as-judge setup.

type and infection status) and then incorporated the metadata into the final generation step; and (2) a few-shot prompting strategy. In Approach (1), the group tested MedGemma (27B Multimodal) for both steps, with confidence thresholds also provided alongside the metadata classifications. In Approach (2), an all-mpnet-base-v2 sentence transformer was used to encode training instances for semantic similarity search. The group experimented with retrieving between 5 and 25 samples as few-shot examples, and tested both the InternVL3-38B and MedGemma-27B models.

Unlike the participating teams, the baselines used a zero-shot approach. Both Gemini and GPT-4o, very large general-purpose models (hundreds of billions of parameters), achieved state-of-the-art performance on both English and Chinese. Meanwhile, the smaller Qwen-VL model (7B) showed performance comparable to the RAG approaches on the English subset, despite no additional prompting. For the Chinese dataset, Qwen-VL also produced competitive results relative to the system submissions.

5 Discussion

Results show that out-of-the-box performance from very large general multimodal models such as Gemini and GPT-4o is highly competitive. That said, the smaller Qwen-VL (7B) model achieved results comparable to the Llama-17B and MedGemma-27B system models that used dedicated RAG few-shot examples in both languages. This suggests that while larger models currently hold a strong advantage, there remains considerable room for optimization and specialization, particularly in identifying models that achieve the best performance-cost trade-off.

In this work, we report both classical NLG evaluation metrics and new LLM-as-judge metrics. The variation in the magnitude of these metrics, and the resulting differences in rankings, suggests that a detailed, comprehensive study against human evaluation is necessary.

Compared to the 2024 MEDIQA-M3G task (Yim et al., 2024) on dermatology, both baseline and system performances were lower. For example, the GPT-4 baseline had a BLEU score of 0.813 and 0.867 BERTScore for English. This may be attributable to differences in response lengths. The dermatology dataset had shorter responses (average

12 words in english, 16 in Chinese for the training set) mostly related to diagnosis, compared to the longer responses in this task (29 English, 43 Chinese) which include care instructions.

6 Conclusion

In this shared task, we found that a range of large multimodal language models, both with and without few-shot examples and RAG augmentations, demonstrated varying levels of performance. While models with larger parameter counts generally held an advantage, performance rankings among models of comparable size were less predictable.

With the highest average score reaching only 0.679, it is clear that further progress is required. Moreover, the variation across evaluation metrics highlights the need for significant advancements in evaluation methodology to reduce the uncertainty associated with individual metrics.

This shared task focused solely on free-text response generation; however, incorporating metadata such as wound type would enable richer multimodal studies in the future. In particular, systems that jointly optimize classification and generation in a fine-tuned setting may yield notable performance gains.

We hope that this work will inspire further research in multimodal patient question answering and medical open-response evaluation, as well as encourage exploration of such applications in clinical practice.

References

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

Bavana Durgapraveen, Sornaraj Sivasankaran, Abhinand Balachandran, and Sriram Rajkumar. 2025. Exl health ai lab at mediqa-wv 2025: Mined prompting and metadata-guided generation for wound care visual question answering. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

A H M Rezaul Karim and Özlem Uzuner. 2025. Masonnlp at mediqa-wv 2025: Multimodal retrieval-augmented generation with large language models for medical vqa. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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

Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M Kennedy Hall, and Meliha Yetisgen. 2025. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024. Overview of the MEDIQA-M3G 2024 shared task on multilingual multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 581–589, Mexico City, Mexico. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.