hgkai26 at MEDIQA-EVAL 2026: Automated Evaluation of Visual Medical Question Answering Using LLM-as-a-Judge

Haritha Gangavarapu
CGI Inc.
harithagangavarapu92@gmail.com

Abstract

As there is a rise in the use of multimodal large language models (LLMs) for medical response generation, it is necessary to have reliable automated evaluation mechanisms that can assess the quality of model-generated outputs. The MediQA-Eval 2026 shared task focuses on grading AI-generated dermatology and wound care responses using structured human-aligned rubrics. In this work, we explore a zero-shot multimodal LLM-as-a-Judge framework to assess candidate responses across multiple quality dimensions. System performance is evaluated using the official task metrics designed to reflect alignment with human judgments. Our findings provide preliminary insights into the feasibility and limitations of LLM-based evaluators for rubric-guided medical response assessment.
Keywords: LLM-as-a-Judge, Multimodal Evaluation, Dermatology and Wound Care QA

1. Introduction

Users often rely on online medical platforms to seek guidance on health-related issues. These platforms typically allow users to submit textual descriptions of their symptoms, and in certain specialties such as dermatology and wound care, accompanying clinical images are provided. Although medical professionals can respond to these questions, the high volume of submissions makes it challenging to provide timely responses.

To address scalability, AI systems can be developed to automatically generate responses based on user queries and associated images. However, automated responses in medical domains may contain inaccuracies, hallucinations, or unsafe recommendations. Since providing accurate information is the highest priority, before such responses can be presented to users, an additional evaluation layer is critical to assess their quality and safety.

The MediQA-Eval 2026 (mediQA26, 2026) evaluation task focuses on developing automated systems capable of grading AI-generated responses in dermatology and wound care settings. Each instance consists of a user query, associated images and multiple candidate responses generated by vision-language models. The objective is to build an evaluation system that assigns structured quality scores aligned with human judgments.

In this work, we implement a zero-shot multimodal LLM-as-a-Judge framework to automatically evaluate candidate responses. As part of the MediQA 2026 task, the hgkai26 team developed an automated evaluation system leveraging GPT-based models to grade three candidate responses from the dermatology and wound care datasets.

2. Related Work

2.1. Medical Multimodal Question Answering

Visual question answering datasets in dermatology and wound care were introduced in prior MEDIQA shared tasks, as part of MEDIQA-M3G (MediQA-M3G, 2024), MEDIQA-WV (MediQA-WV, 2025), and MEDIQA-MAGIC (ImageCLEF24, 2024) and 2025 (ImageCLEF25, 2025). These benchmarks consist of medical user queries along with associated images and have been used to study multimodal medical question answering.

2.2. Evaluation of Natural Language Generation Systems

Evaluating natural language generation (NLG) systems is inherently challenging, particularly in domains requiring domain expertise such as medicine. Traditional automatic metrics such as BLEU and ROUGE rely primarily on lexical overlap with reference texts. While these metrics are useful in structured generation tasks, they may not adequately capture semantic correctness, factual accuracy, or clinical safety in medical question answering(Celikyilmaz et al., 2021). In the evaluation of medical question answering systems, responses must be assessed not only for lexical overlap with reference texts, but also for factual accuracy and clinical safety. As a result, human evaluation using structured rubrics is often considered the gold standard. However, manual evaluation is expensive, time-consuming and difficult to scale. Many prior works have explored the wide use of large language models as automated evaluators (LLM-as-a-Judge) for NLG tasks (Liu et al., 2023; Xu, 2023; Saha et al., 2025).

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 257–261

Split	Dataset	Queries	Systems	Responses	Metrics	Reviewers	Total Rows
Valid	iiyi	56	3	168	6	2	2016
Valid	woundcare	105	3	315	6	1	1890
Test	iiyi	100	3	300	6	2	3600
Test	woundcare	93	3	279	6	2	3348

Table 1: Number of instances by split for the English subset of the task dataset. The total number of rows corresponds to the product of the number of system responses, evaluation metrics, and reviewers.

3. Data

3.1. Task Dataset

The MediQA 2026 evaluation task builds upon existing dermatology and wound care visual question answering benchmarks. The dermatology subset consists of user queries originally posted on iiyi, a Chinese online medical consultation platform that allows users to ask their questions related to medical conditions or concerns and are answered by medical professionals. The wound care subset includes user queries and associated images collected from tieba.baidu.com and zhidao.baidu.com (wai Yim et al, 2025). Each instance includes three system-generated responses produced by Gemini-1.5-Pro, GPT-4o, and Qwen-VL-Chat using the original query and associated images as input context.

The dataset includes English (EN) and Chinese (ZH) subsets. The validation split contains 168 (iiyi) and 315 (wound care) English instances evaluated across six metrics, with corresponding Chinese subsets evaluated across two metrics, resulting in 3,864 human evaluation instances.

The test split contains 300 (iiyi) and 279 (wound care) English instances evaluated across six metrics, with corresponding Chinese subsets evaluated across two metrics, resulting in 4,632 human evaluation instances overall. Table 1 summarizes the number of annotation instances by split for the English subset.

In this work, we focus exclusively on the English evaluation track. To comply with task submission requirements, predictions for the Chinese (ZH) subset were assigned a placeholder value of -1 in our submission.

3.2. Annotator agreement

Given the subjective nature of response quality evaluation, we analyzed the inter-annotator agreement of the provided human annotations on the English test subset, where each instance was graded independently by two human reviewers. Table 2 reports agreement rates across evaluation metrics and datasets.

Inter-annotator agreement was computed locally by comparing the scores assigned by the two reviewers for each metric. A binary agreement score was assigned - 1 if both reviewers assigned the same rubric value, or 0 otherwise. Although the scoring rubric uses numeric values (e.g., {0, 0.5, 1}), these were treated as categorical labels for the purpose of agreement measurement.

As shown in Table 2, aggregate agreement differed substantially across datasets. The wound care subset exhibited higher *overall* agreement (81%) compared to the iiyi dermatology subset (56%).

For the iiyi dataset, writing style demonstrated comparatively lower agreement across systems. Metrics following the three-level rubric ({0, 0.5, 1}) showed lower agreement than the binary disagreement flag. In contrast, for the wound care dataset, writing style and relevance showed relatively higher agreement score. It is important to note that the reviewer groups for the iiyi and wound care datasets were not the same, which may partially contribute to the observed differences in agreement.

System	Metric	iiyi (n=100)	woundcare (n=93)
SYSTEM001			
	Completeness	0.48	0.87
	Disagree_flag	0.82	0.84
	Factual_accuracy	0.59	0.78
	Overall	0.57	0.74
	Relevance	0.66	0.99
	Writing_style	0.37	0.94
SYSTEM002			
	Completeness	0.48	0.87
	Disagree_flag	0.86	0.84
	Factual_accuracy	0.63	0.73
	Overall	0.66	0.69
	Relevance	0.69	0.98
	Writing_style	0.35	0.96
SYSTEM003			
	Completeness	0.30	0.62
	Disagree_flag	0.65	0.75
	Factual_accuracy	0.56	0.62
	Overall	0.55	0.59
	Relevance	0.61	0.85
	Writing_style	0.32	0.85

Table 2: Inter-annotator agreement across evaluation metrics and systems on the English test subset

4. Methods

4.1. Evaluation criteria

Each system response was evaluated across four dimensions: *completeness*, *factual accuracy*, *relevance*, and *writing style*. In addition, an *overall* score was assigned. Scores for these dimensions were given on a rubric of {0, 0.5, 1}, where 1 indicates correct, accurate, and fully relevant, 0.5 indicates a partially correct or incomplete response, and 0 indicates an incorrect or irrelevant response.

Additionally, a binary *disagree-flag* score was assigned, where 1 indicates disagreement with the candidate response and 0 indicates agreement.

These structured scores provide an overall assessment of the quality and reliability of the system response and help determine whether it is suitable to be presented to users.

4.2. GPT-as-a-Judge evaluation framework

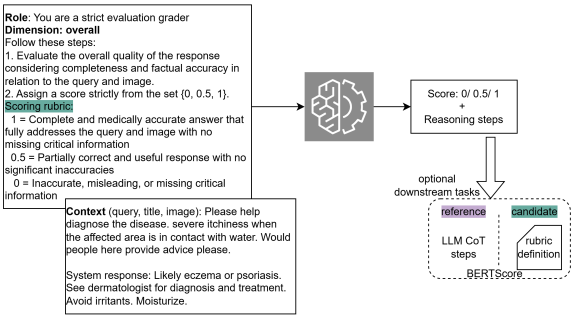

Figure 1: Example of a rubric-based instruction used for response evaluation.

We use a multimodal LLM-as-a-Judge setup based on OpenAI's `score_model` evaluation framework (OpenAIEvals, 2023). Each system response is graded across six metrics: *completeness, factual accuracy, relevance, writing style, overall, and disagreement flag*. For each instance, the grader receives the user query title (EN), query text (EN), associated image, and the candidate system response.

The model is instructed to act as a strict medical grader and assign a score from the set {0, 0.5, 1} according to rubric definitions provided in the prompt shown in Figure 1. As this is a zero-shot setup, no examples are provided to illustrate how the rubric levels 0, 0.5, and 1 should differ. The model relies solely on the metric specific rubric guidance when assigning scores.

We evaluated multiple GPT variants(OpenAI, 2024), including GPT-4o-mini and GPT-5-mini, under different parameter settings, including deterministic setting (temperature = 0). While our official submission included only the GPT-4o-mini run, we conducted additional experiments using GPT-5-mini after the test submission window closed. These additional evaluations were run locally using the official evaluation script.

In addition to numeric scores, the model also generates intermediate reasoning. We conducted preliminary analyses of these reasoning outputs to explore whether additional alignment signals could be extracted. However, as the observed performance was not sufficiently strong, these signals were not incorporated into the final system and are left for future work.

5. Results

We report results across the two datasets, iiyi and wound care, focusing on the *overall* metric as well as results computed on the combined data. To remain consistent with the task evaluation, we report the aggregated correlation metrics in Table 3. Additionally, in Table 4 we report classification accuracy for the *disagreement* dimension. Although accuracy was not part of the official evaluation metrics, we include it to quantify exact score agreement.

Kendall and Spearman measure rank-based correlation, while Pearson measures linear correlation. These metrics are appropriate for ordinal scoring settings, as they assess alignment between the system's score ordering and human judgments. In contrast, accuracy treats scores as discrete class labels (0, 0.5, 1) and measures exact agreement between system and human scores.

- ALL-en-overall-mean: The average of Kendall, Spearman, and Pearson correlations computed over all English data rows for the *overall* metric.
- ALL-en-ALL-mean: The average of Kendall, Spearman, and Pearson correlations computed over all English data rows and averaged across all evaluation metrics.

5.1. Observations on Zero-Shot Grading

While the overall correlation with human scores remained low across all configurations including stricter rubric prompts and the use of a larger model(GPT-4o), GPT-5-mini achieved higher correlation compared to the other setups. GPT-5-mini ranked weak system responses more similarly to the human reviewers. However, it assigned more conservative scores, particularly for the *overall* metric. In many cases where other models predicted a score of 1, GPT-5-mini predicted 0 or 0.5 instead. As a result, while classification accuracy remained low when treating scores as discrete labels, rank-based correlation metrics were relatively stronger.

Model	Params	Validation		Test	
		ALL-en-overall-mean	ALL-en-ALL-mean	ALL-en-overall-mean	ALL-en-ALL-mean
GPT-5-mini	default	0.3399	0.3095	0.3059	0.2547
GPT-4o-mini*	default	0.2811	0.2816	0.2444	0.2129
GPT-4o-mini	temperature=0	0.2845	0.3204	0.1757	0.2101

Table 3: Mean correlation results on the validation and test sets across all metrics and the Overall dimension. * indicates the official system submission

Model	Params	Disagree-flag Accuracy		Disagree-flag Mean Correlation	
		Woundcare	iiyi	Woundcare	iiyi
GPT-5-mini	default	**0.79**	**0.73**	**0.2251**	**0.2686**
GPT-4o-mini*	default	0.78	0.70	0.1866	0.2018
GPT-4o-mini	temperature=0	0.68	0.61	0.0902	0.1981

Table 4: Disagree-flag accuracy and mean correlation across datasets in the test split. * indicates the official system submission

We also observed that the deterministic setup (temperature = 0) achieved better correlation scores on the validation set. However, this improvement did not generalize to the test set, suggesting that deterministic setting alone does not consistently improve alignment with human evaluations. These findings suggest that the limitation lies not primarily in parameters or model scale, but in the zero-shot direct scoring setup itself. In this setting, the LLM judge must interpret the rubric definitions without any calibration examples. Which could be challenging to distinguish between fine-grained score levels, especially in medical evaluation tasks.

6. Conclusion

Our system explores a zero-shot rubric-based grading approach using the `score_model` evaluation type of the OpenAI evaluation framework. When the grading prompt included explicit structural constraints or deterministic settings, performance declined on the iiyi dataset, which also shows lower annotator agreement. Additionally, the accuracy for wound care with and without deterministic configuration (temperature = 0) was 0.78 vs 0.72, while for iiyi it was 0.49 vs 0.46, indicating lower agreement overall. Deterministic grading does not necessarily improve correlation with human judgments in multimodal medical evaluation. As a result, scoring occasionally exhibited variability across rubric boundaries. In future work, incorporating concrete examples to better distinguish rubric score levels and averaging across multiple runs may improve the stability of the evaluations. For the *overall* metric, GPT-5-mini provided more restrictive ratings compared to other setups, suggesting that restrictive scoring behavior may contribute to improved rank correlation.

7. Bibliographical References

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey.

ImageCLEF24. 2024. Imageclef 2024 medical: Mediqa. https://www.imageclef.org/2024/medical/mediqa. Accessed: 2026-02-18.

ImageCLEF25. 2025. Imageclef 2025 medical: Mediqa. https://www.imageclef.org/2025/medical/mediqa. Accessed: 2026-02-18.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

MediQA-M3G. 2024. Mediqa-m3g shared task. https://sites.google.com/view/mediqa2024/mediqa-m3g. Accessed: 2026-02-18.

MediQA-WV. 2025. Mediqa-wv shared task. https://sites.google.com/view/mediqa-2025/mediqa-wv. Accessed: 2026-02-18.

mediQA26. 2026. Mediqa-eval 2026. https://sites.google.com/view/mediqa2026/mediqa-eval. Accessed: 2026-02-18.

OpenAI. 2024. Gpt-4 technical report.

OpenAIEvals. 2023. Openai evals framework. https://github.com/openai/evals. Accessed: 2026-02-18.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan reason for evaluation with thinking-llm-as-a-judge.

Wen wai Yim et al. 2025. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888.

Wenda et al. Xu. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.