Overview of the MEDIQA-EVAL 2026 Shared Task on Evaluation Metrics in Medical Multimodal Question Answering

Asma Ben Abacha, Wen-wai Yim
{abenabacha, yimwenwai}@microsoft.com
Microsoft, Health and Life Sciences AI, Redmond, US

Abstract

Evaluating clinical text generation remains challenging, as automatic metrics often correlate weakly with clinician judgments. This issue is particularly pronounced in medical multimodal question answering (MMQA), where systems must integrate visual and textual information and evaluation must capture factual accuracy, visual grounding, completeness, and overall coherence. Despite rapid progress in MMQA, there is limited consensus on clinically meaningful evaluation, and existing metrics, largely adapted from general NLG or VQA, often fail to capture domain-specific criteria. We introduce MEDIQA-EVAL 2026, a shared task on evaluation metrics for medical multimodal QA. To our knowledge, this is the first shared task focused on evaluating automatic metrics in this setting. We release a dataset of medical visual question-answer pairs annotated with multidimensional clinician judgments. Systems are evaluated by the correlation of their metric scores with expert ratings on a held-out test set. Participants explored diverse approaches, including vision-language models, retrieval-augmented judging, metric-specific classifiers, reinforcement learning, and LLM-as-a-judge frameworks. Results show that model-based evaluators achieve stronger alignment with human judgments than traditional NLG metrics, particularly on English data, while performance remains lower on Chinese, highlighting challenges in multilingual evaluation. Notably, our MEDIQA LLM-as-a-judge approach achieves strong performance across both languages.

Keywords: Evaluation, Medical Multimodal Question Answering, LLM-as-a-Judge, Human Alignment

1. Introduction

Evaluating clinical text generation remains challenging, particularly in medical settings where assessments may vary across experts and correctness is often graded rather than strictly binary. Widely used reference-based automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), and BLEURT (Sellam et al., 2020) have been shown to correlate weakly with clinician judgments in clinical contexts. Studies on automated medical note generation further demonstrates that standard metrics do not reliably capture clinically important properties such as factual accuracy, omissions, and hallucinations, and that domain-adapted or composite metrics may better approximate expert assessment (Ben Abacha et al., 2023).

LLM-based evaluators such as GPTScore (Fu et al., 2024) have been proposed to assess generated text along quality dimensions, offering greater flexibility for open-ended generation tasks. More recently, the MORQA benchmark for medical open-response QA finds that LLM-as-a-judge approaches (e.g., GPT-4 and Gemini) tend to correlate more strongly with expert ratings than traditional metrics across multilingual, multimodal QA datasets, in part because they are more sensitive to semantic nuances and robust to variability among multiple valid references (Yim et al., 2026). These insights highlight the need for evaluation methods that go beyond surface overlap and better align with clinical judgment in complex QA scenarios.

The challenge is further amplified by advances in Vision–Language Models (VLMs). Although VLMs can generate fluent, human-like responses, reliably evaluating the accuracy and completeness of their outputs remains difficult, particularly in multimodal settings where images must be interpreted alongside textual information and where multiple clinically acceptable responses may exist. In medical multimodal question answering (MMQA), systems must integrate visual evidence and clinical language, yet evaluation practices largely remain adapted from general natural language generation and VQA. While datasets such as VQA-RAD (Lau et al., 2018), VQA-Med (Ben Abacha et al., 2019, 2020, 2021), and PathVQA (He et al., 2020) have enabled rapid model development, evaluation has often relied on exact-match or classification-style accuracy for short, constrained answers. Such setups only partially reflect clinical needs in open-ended MMQA.

To address this gap, we introduce the MEDIQA-EVAL 2026 shared task on Evaluation Metrics for Medical Multimodal Question Answering. We extend the MEDIQA-M3G 2024, MEDIQA-MAGIC 2024, MEDIQA-WV 2025, and MEDIQA-MAGIC-2025 (Yim et al., 2024a,b, 2025c,a) shared tasks on dermatology and wound-care visual question answering by shifting the focus to evaluating open-ended system responses. In this multimodal evaluation task, participants develop automatic metrics that assign quality scores to model-generated answers for patient questions paired with one or multiple clinical images. Responses are assessed

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 1–11
16 May 2026.

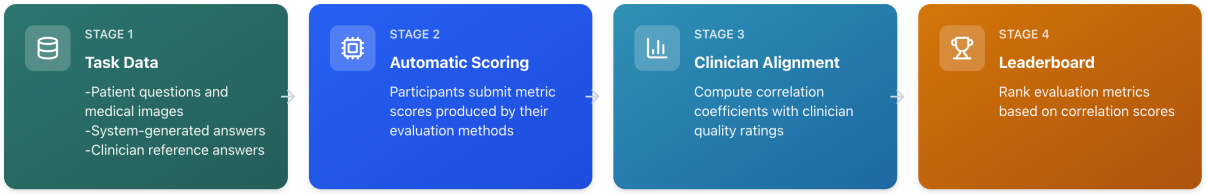

Figure 1: Overview of the meta-evaluation pipeline used in the MEDIQA-EVAL 2026 shared task. Stage 1 provides the multimodal dataset with patients' questions and medical answers, system-generated answers, and clinician reference answers. In Stage 2, participants submit automatic metric scores computed over the provided system-generated answers. Stage 3 evaluates metric validity by measuring correlation between metric scores and clinician quality ratings. Stage 4 ranks submitted metrics on a leaderboard based on their alignment with clinician judgments.

across multiple clinically motivated quality dimensions, including completeness, factual accuracy, relevance, writing style, and overall quality.

We provide clinician ratings for the development set and rank submitted metrics based on their correlation with expert judgments on the test set. By centering alignment with clinician assessments and adopting a multidimensional evaluation framework, this shared task establishes a benchmark for MMQA metric evaluation and promotes the development of reliable, clinically grounded evaluation methods for multimodal clinical QA systems.

2. Task Description

We introduce a clinician-aligned meta-evaluation framework for assessing automatic evaluation metrics in medical question answering. Unlike conventional shared tasks that rank model outputs, the MEDIQA-EVAL shared task[1] evaluates the quality of evaluation metrics themselves. The primary objective is to determine the extent to which automatic metrics align with expert clinician judgments. Figure 1 illustrates the overall evaluation pipeline. Rather than submitting system predictions, participants submit metric-generated scores, which are subsequently evaluated based on their agreement with clinician ratings.

2.1. Stage 1 - Task Data

The MEDIQA-EVAL 2026 dataset[2] spans two languages (English and Chinese) and consists of:

- Patient questions and associated medical images,
- System-generated answers produced by multimodal medical QA systems, and associated clinician quality ratings (provided in the development set only),
- Clinician reference answers.

Clinician ratings serve as the gold standard for metric validation. These annotations are multidimensional and language-specific. For the English datasets, responses are evaluated across several quality dimensions, including completeness, factual accuracy, relevance, writing style, and overall quality. For the Chinese datasets, domain experts assess factual consistency with the gold standard and writing style.

2.2. Stage 2 - Response Scoring

Participants are provided with the dataset described above and are required to compute quality scores using their proposed evaluation metric. For each system-generated answer, their proposed metric must output a score reflecting the estimated response quality.

Participants do not submit model outputs. Instead, they submit metric-generated scores computed over the provided system answers to ensure that all methods are evaluated on the same inputs.

2.3. Stage 3 - Clinician Alignment

To assess metric quality, we measure the statistical correlation between participant-submitted metric scores and clinician quality ratings (Figure 1, Stage 3). Because clinician evaluation spans multiple quality dimensions, we compute correlations both at the dimension level and for overall quality:

- For the EN datasets, clinician ratings include the following dimensions: *disagreement flag*, *completeness*, *factual accuracy*, *relevance*, *writing style*, and *overall quality*.
- For the ZH datasets, ratings include *factual consistency with the gold standard* and *writing style*.

[1] https://sites.google.com/view/mediqa2026/mediqa-eval

[2] The MEDIQA-EVAL 2026 dataset is available at https://osf.io/kcv2n

Let m_i denote the score assigned by a metric to instance i, and let $c_i^{(d)}$ denote the clinician rating for instance i on dimension d (e.g., completeness, factual accuracy, writing style, or overall quality). Metric validity with respect to a given dimension d is quantified as:

$$\text{Alignment}_d(m) = \text{corr}(m_i, c_i^{(d)}) \quad (1)$$

In addition, we report alignment with the overall clinician rating when available. Correlation is computed using **Pearson's** r, **Spearman's** ρ, and **Kendall's** τ. Higher correlation indicates stronger agreement with clinician judgment on the evaluated dimension and, therefore, greater metric reliability.

2.4. Stage 4 - Leaderboard & Results

Metrics are ranked according to their correlation scores (Figure 1, Stage 4). Specifically, leaderboard positions are determined by the strength of alignment between metric scores and expert annotations across the evaluated quality dimensions. The competition therefore identifies evaluation methods that most faithfully approximate expert clinical judgment. By grounding metric validation in multidimensional clinician assessment, MEDIQA-EVAL moves beyond surface-level similarity measures and prioritizes evaluation approaches that better capture clinical relevance, factual accuracy, completeness, and safety-related considerations.

3. Datasets & Annotations

We use the MORQA benchmark (Yim et al., 2026), which is built upon two dermatology-focused visual question answering (VQA) datasets: WoundcareVQA (Yim et al., 2025b) and DermaVQA-iiyi (Yim et al., 2024c). Both source datasets contain English and Chinese question-answer pairs and are centered on dermatology, a visually grounded medical specialty in which clinical assessment heavily depends on images. Each instance consists of a patient question paired with corresponding medical image(s) and clinician-provided reference answers.

In the MORQA benchmark[3], we further augment this data with system-generated answers from multimodal medical QA models and associated expert quality ratings. These additional annotations enable direct evaluation of automatic metrics by measuring their alignment with clinician judgments.

3.1. English Datasets

For the English (WoundcareVQA and DermaVQA-iiyi) datasets, each system-generated response was evaluated by a practicing medical doctor using a multidimensional annotation framework. Annotations include a binary disagreement flag (`disagree_flag`) indicating whether the expert disagreed with the response, as well as graded assessments of `completeness`, `factual_accuracy`, `relevance`, `writing_style`, and `overall`.

All graded dimensions are scored on a three-point scale $\{0, 0.5, 1.0\}$, where 1.0 denotes a fully satisfactory response, 0.5 denotes a partially satisfactory response, and 0 denotes an inaccurate, incomplete, or otherwise inadequate response.

3.2. Chinese Datasets

For the Chinese (WoundcareVQA and DermaVQA-iiyi) datasets, each system-generated response was evaluated by a single domain expert trained at a Chinese medical school. The annotation framework includes two dimensions: `factual_consistency_wgold` and `writing_style`.

`factual_consistency_wgold` measures the degree to which a system response is factually consistent with the clinician-provided reference (gold) answer, while `writing_style` assesses the appropriateness and clarity of the response. Both dimensions are scored on a three-point scale $\{0, 0.5, 1.0\}$, where 1.0 indicates full satisfaction, 0.5 indicates partial satisfaction, and 0 indicates an incorrect, inconsistent, or inappropriate response.

3.3. Evaluation Metrics (English)

Following Section 2.3 (Stage 3 - Clinician Alignment), all submitted metrics are evaluated by measuring their correlation with clinician ratings under the evaluation script provided by the task[4]. Correlation is computed using Kendall's τ, Pearson's r, and Spearman's ρ.

For the English datasets, alignment is computed separately for each annotated quality dimension, reflecting the multidimensional clinician evaluation framework. In total, 73 metrics are reported, defined by:

- **Dataset**: {iiyi, woundcare, ALL}
- **Evaluation Dimension (EN_METRIC)**: {en-disagree_flag, en-completeness, en-factual-accuracy, en-relevance, en-writing-style, en-overall}

For each Dataset–EN_METRIC pair, the following correlation scores are computed:

1. Dataset-EN_METRIC-kendalltau

[3] https://osf.io/kcv2n

[4] https://github.com/wyim/MEDIQA-EVAL-2026

2. Dataset-EN_METRIC-pearson
3. Dataset-EN_METRIC-spearman
4. Dataset-EN_METRIC-mean

where `mean` denotes the average of the three correlation coefficients. In addition, we report an aggregated score (ALL-en-ALL-mean), computed by averaging correlations across all English datasets and evaluation dimensions.

3.4. Evaluation Metrics (Chinese)

For the Chinese datasets, metric validity is likewise assessed via correlation with clinician annotations, as described in Stage 3. Given the language-specific annotation framework, alignment is computed with respect to two quality dimensions. In total, 25 metrics are reported, defined by:

- **Dataset**: {iiyi, woundcare, ALL}
- **Evaluation Dimension (ZH_METRIC)**: {zh-factual-consistency-wgold, zh-writing-style}

For each Dataset–ZH_METRIC pair, the following correlation scores are computed:

1. Dataset-ZH_METRIC-kendalltau
2. Dataset-ZH_METRIC-pearson
3. Dataset-ZH_METRIC-spearman
4. Dataset-ZH_METRIC-mean

where `mean` denotes the average of the three correlation coefficients. We additionally report an overall aggregated score (ALL-zh-ALL-mean), obtained by averaging correlations across both Chinese datasets and evaluation dimensions.

4. Baseline Systems

4.1. Overview

We evaluate candidate medical responses using a Large Language Model (LLM)-as-a-Judge framework. The judge receives as input: (1) the original patient query, (2) all associated medical images, (3) a set of expert-written reference responses, and (4) a candidate response to evaluate. The model outputs metric-specific scores aligned with the human evaluation protocol.

We investigate both a single-judge configuration and multi-judge ensemble variants.

4.2. MEDIQA LLM-as-a-Judge

4.2.1. Prompt Design

Full prompts are provided in Appendix.

For English samples, the judge outputs:

- `disagree_flag` $\in \{0, 1\}$
- `completeness` $\in [0.00, 1.00]$
- `factual-accuracy` $\in [0.00, 1.00]$
- `relevance` $\in [0.00, 1.00]$
- `writing-style` $\in [0.00, 1.00]$
- `overall` $\in [0.00, 1.00]$

For Chinese samples, the judge outputs:

- `factual-consistency-wgold` $\in [0.00, 1.00]$
- `writing-style` $\in [0.00, 1.00]$

The prompt includes:

- Explicit metric definitions.
- Calibration anchors (e.g., ≥ 0.85 for strong agreement, ≤ 0.30 for contradiction).
- Constraints enforcing metric independence (e.g., `overall` cannot exceed both `completeness` and `factual-accuracy`).
- Instructions to use reference responses and images as primary evidence.

4.2.2. Reference Handling

All expert reference responses are provided to the model. References are sorted by (1) frequency indicator and (2) completeness score, so that more reliable responses appear earlier. This ordering helps the judge prioritize consensus medical conclusions.

4.2.3. Continuous Scoring and Snapping

The model outputs continuous scores in $[0.00, 1.00]$. To stabilize ranking behavior and reduce numeric noise, we apply post-processing snapping to 0.05 increments:

$$\hat{s} = \text{round}\left(\frac{s}{0.05}\right) \times 0.05$$

4.2.4. Image Integration

All available images are included in the prompt as visual inputs before textual content. This ensures the judge conditions its reasoning on visual evidence, which is particularly important for woundcare cases.

4.3. Multi-Judge Ensembles

4.3.1. Temperature Ensemble

We run the LLM three times using different temperature settings: $T \in \{0.0, 0.3, 0.7\}$.

For each metric:

- Continuous scores are aggregated using the median.
- `disagree_flag` is aggregated via majority vote.

4.3.2. Persona Ensemble

We design three specialized judge prompts:

1. Diagnostic Consistency Judge
2. Clinical Safety Judge
3. Professional Communication Judge

Persona diversity encourages complementary reasoning patterns. Each judge independently produces metric scores using the same output schema. Aggregation follows the same median (continuous) and majority vote (binary) strategy.

4.4. Implementation Details

We use GPT-4o as the base model. Decoding is performed deterministically for the single-judge setting ($T = 0$). Outputs are constrained using strict JSON schema validation with automatic retry upon failure. No fine-tuning or in-context learning is applied.

5. Participating Teams & Results

The MEDIQA-EVAL 2026 shared task attracted 48 registered teams from academy and industry. Among them, five finalist teams submitted their codes and runs following the challenge rules[5]. Table 1 presents the teams that participated in the two subtasks (English and Chinese). We limited the number of submitted runs to 15 runs per team.

5.1. Code Verification

For additional validation, we required the submission of the code in addition to the models' outputs/runs. The participants shared their private codes with the organizers on GitHub following provided guidelines.

5.2. Official Results

The main results are summarized[6] in Tables 2 and 3. Our baseline systems were not included in the official ranking. On the English dataset, the SUAT-BMI team achieved the highest ALL-en-ALL-mean score (0.4816), followed by our single-LLM baseline (0.4524). On the Chinese dataset, the MedAware team obtained the highest official ALL-zh-ALL-mean score (0.2940), while our single-LLM baseline, although not part of the official ranking, achieved a higher score of 0.3271.

Our MEDIQA baseline (single-LLM) achieves strong performance on both datasets. While ensemble variants introduce additional diversity, they do not outperform the single-LLM baseline in terms of ALL-en-ALL-mean or ALL-zh-ALL-mean, suggesting that increasing output diversity alone is insufficient to improve ranking performance. Specifically, the temperature ensemble achieves ALL-en-ALL-mean and ALL-zh-ALL-mean scores of 0.4507 and 0.2767, respectively, while the persona ensemble achieves 0.4519 and 0.3032. These findings highlight the robustness of the single-LLM approach and suggest that model consistency plays a more important role than diversity in this task.

Table 4 and Table 5 present detailed results for all English (EN) metrics and correlation scores. Table 6 provides detailed results for all Chinese (ZH) metrics. Beyond overall performance, several trends emerge from the detailed results. First, performance varies across evaluation dimensions, with higher correlations observed for factual consistency and relevance, and lower scores for disagreement detection and writing quality, indicating that stylistic and contradiction-related aspects remain challenging. Second, differences are observed across correlation metrics, with Kendall's τ often yielding lower values than Pearson and Spearman correlations, reflecting its greater sensitivity to pairwise ranking differences. Despite these differences, the correlation metrics exhibit consistent trends across systems, indicating strong agreement in system rankings and reinforcing the reliability of the evaluation. Third, performance on the Chinese dataset is generally lower than on the English dataset, highlighting challenges in multilingual generalization.

Table 7 presents the correlation between common NLG metrics (ROUGE-1, BERTScore, and BLEURT, computed as both the maximum and mean across multiple references) and human judgments, alongside the MEDIQA LLM-as-a-judge baseline. On the English dataset, MEDIQA LLM-as-a-judge achieves the highest agreement with human overall ratings across all correlation measures, outperforming both traditional NLG metrics and participating systems (cf. Table 5). On the Chinese dataset, BLEURT (mean) performs best overall, with BERTScore (mean) remaining competitive, particularly for Spearman correlation. These results highlight the advantage of learned evaluators over surface-level metrics, especially in multilingual settings.

[5] Leaderboard: https://www.codabench.org/competitions/12115

[6] Full results are available at https://tinyurl.com/mediqaevalres

	Team	Affiliation	Subtasks	Paper	Code
1	SUAT-BMI	Shenzhen University of Advanced Technology, China	EN	(Peng et al., 2026)	[1]
2	SloCal-Net	University of Maribor, Slovenia & Magna Graecia University, Italy	EN/ZH	(Kocbek et al., 2026)	[2]
3	MedAware	McGill University, Canada	EN/ZH	(Hao and Liu, 2026)	[3]
4	BDI	University of Oxford, UK & University of California, USA	EN/ZH	(Xu et al., 2026)	[4]
5	hgkai26	CGI Inc., USA	EN	(Gangavarapu, 2026)	[5]

[1] https://github.com/newchat111/MEDIQA-MEDGEMMA
[2] https://github.com/pkocbek/MEDIQA-EVAL26
[3] https://github.com/Ziqi-Hao/MedAware-for-mediqa-2026
[4] https://github.com/justin13601/mediqa-eval-bdi
[5] https://github.com/harithagmu/MediQAEval2026

Table 1: MEDIQA-EVAL 2026: Participating teams, subtasks, papers, and codes.

Team	Run ID	ALL-en-ALL-mean	ALL-zh-ALL-mean
SUAT-BMI	526661	0.48156203	-
SloCal-Net	521002	0.41575853	0.25546085
MedAware	526509	0.39093232	0.25053383
BDI	526670	0.36918917	0.25244488
hgkai26	527947	0.21286092	-
MEDIQA Organizers	*526144*	*0.45239557*	*0.26694508*

Table 2: Official Results: Best Runs on the English Dataset (and their results on the Chinese Dataset)

Team	Run ID	ALL-zh-ALL-mean	ALL-en-ALL-mean
MedAware	524265	0.29400372	0.38780292
BDI	523374	0.2686269	0.33431094
SloCal-Net	524450	0.25956562	0.39715838
MEDIQA Organizers	*526255*	*0.32714034*	*0.44388561*

Table 3: Official Results: Best Runs on the Chinese Dataset (and their results on the English Dataset)

Team	Run ID	Disagree				Completeness				Factual			
		K	P	S	M	K	P	S	M	K	P	S	M
SUAT-BMI	526661	0.349	0.372	0.372	0.364	**0.416**	0.486	0.459	**0.454**	0.485	**0.553**	**0.574**	**0.537**
SloCal-Net	521002	0.250	0.261	0.266	0.259	0.392	0.423	0.409	0.408	0.475	0.523	0.518	0.505
MedAware	526509	0.117	0.118	0.118	0.118	0.316	0.397	0.336	0.350	**0.511**	0.528	0.551	0.530
BDI	526670	0.384	0.385	0.384	0.385	0.311	0.351	0.324	0.329	0.403	0.423	0.431	0.419
hgkai26	527947	0.203	0.203	0.203	0.203	0.149	0.162	0.155	0.155	0.178	0.208	0.189	0.191
MEDIQA	526144	**0.393**	**0.392**	**0.394**	**0.393**	0.398	**0.493**	**0.460**	0.450	0.414	0.504	0.483	0.467

Table 4: Official Results: Best Runs on the English Dataset (EN metrics - Part 1). K = Kendall's τ, P = Pearson correlation, S = Spearman correlation, and M = mean correlation.

Team	Run ID	Relevance				Writing				Overall				ALL
		K	P	S	M	K	P	S	M	K	P	S	M	Mean
SUAT-BMI	526661	**0.587**	**0.571**	**0.627**	**0.595**	**0.380**	0.403	**0.399**	**0.394**	0.482	0.579	0.574	0.545	**0.482**
SloCal-Net	521002	0.481	0.478	0.504	0.488	0.335	0.351	0.348	0.345	0.462	0.510	0.497	0.490	0.416
MedAware	526509	0.463	0.459	0.495	0.472	0.322	0.382	0.345	0.350	**0.492**	0.546	0.540	0.526	0.391
BDI	526670	0.346	0.317	0.357	0.340	0.293	0.316	0.301	0.304	0.418	0.456	0.444	0.439	0.369
hgkai26	527947	0.261	0.293	0.271	0.275	0.199	0.220	0.205	0.208	0.231	0.258	0.244	0.244	0.213
MEDIQA	526144	0.415	0.492	0.488	0.465	0.339	**0.446**	0.380	0.388	0.476	**0.592**	**0.583**	**0.551**	0.452

Table 5: Official Results: Best Runs on the English Dataset (EN metrics - Part 2). K = Kendall's τ, P = Pearson correlation, S = Spearman correlation, and M = mean correlation.

Team	Run ID	Factual Consistency				Writing Style				Overall
		Kendall	Pearson	Spearman	Mean	Kendall	Pearson	Spearman	Mean	
MedAware	524265	**0.471**	**0.511**	**0.511**	**0.498**	0.105	0.062	0.105	0.090	0.294
BDI	523374	0.330	0.365	0.366	0.354	0.183	0.183	0.183	0.183	0.269
SloCal-Net	524450	0.412	0.437	0.438	0.429	0.091	0.088	0.092	0.090	0.260
MEDIQA	526255	0.386	0.464	0.460	0.437	**0.193**	**0.244**	**0.215**	**0.217**	**0.327**

Table 6: Official Results: Best Runs on the Chinese Dataset (all ZH metrics)

NLG Metric	English Dataset - Overall Metric				Chinese Dataset - Factual Consistency			
	Kendall	Pearson	Spearman	Mean	Kendall	Pearson	Spearman	Mean
ROUGE-1 (max)	0.351	0.439	0.463	0.418	0.002	-0.022	0.002	-0.006
ROUGE-1 (mean)	0.372	0.501	0.498	0.457	0.002	-0.016	0.002	-0.004
BERTScore (max)	0.296	0.400	0.395	0.364	0.325	0.409	0.413	0.383
BERTScore (mean)	0.356	0.490	0.476	0.441	0.376	0.473	**0.477**	0.442
BLEURT (max)	0.333	0.439	0.440	0.404	0.336	0.425	0.427	0.396
BLEURT (mean)	0.378	0.523	0.497	0.466	0.378	**0.480**	0.476	**0.445**
MEDIQA LLM-as-a-judge	**0.476**	**0.592**	**0.583**	**0.551**	**0.386**	0.464	0.460	0.437

Table 7: Comparison of NLG metrics and LLM-as-a-judge against human evaluations on English (overall quality) and Chinese (factual consistency) datasets. MEDIQA LLM-as-a-judge achieves the strongest alignment with human overall ratings in English, while BLEURT (mean) performs best in Chinese.

6. System Descriptions

6.1. SUAT-BMI Team

The SUAT-BMI team used an ensemble evaluation framework that combines (i) few-shot LLM judging, (ii) metric-specific discriminative classifiers, and (iii) retrieval-augmented prompting. First, Qwen 30B-A3B was employed as a judge LLM in a few-shot setting, where the prompt instructs the model to compare an LLM-generated diagnosis/treatment response against multiple physician-written gold responses treated as ground truth. The judge was calibrated with labeled exemplars and then applied to rate a target sample consisting of the patient query, the candidate response, and the gold responses. Second, PubMedBERT was fine-tuned as a lightweight automatic rater. Prior to fine-tuning, MedGemma was used for zero-shot medical image captioning; the resulting caption was concatenated with the patient query and the candidate response to form an information-rich sequence. Separate PubMedBERT classifiers were trained for each evaluation metric, framing scoring as a three-way classification over 0, 0.5, 1.0, resulting in six distinct metric-specific models. Third, to mitigate hallucinations and improve evidence grounding, retrieval-augmented generation (RAG) was incorporated into the few-shot judging process. A two-stage retrieval pipeline based on LlamaIndex was used to retrieve clinical guideline passages relevant to each gold response, which were appended to the judge prompt under a dedicated "Relevant Clinical Guidelines" section.

In the final ensemble, the fine-tuned PubMedBERT was used for writing-style, RAG-enhanced few-shot judging for relevance, and standard few-shot judging for disagree-flag, factual accuracy, overall, and completeness. The system achieved an official leaderboard score of 0.481.

6.2. SloCal-Net Team

The SloCal-Net team adopted a unified pipeline consisting of evidence retrieval, context grounding, reasoning, and evaluation, with GPT-5-mini serving as the primary reasoning component. In the retrieval stage, ChatGPT Deep Research was used to automatically generate and rank 25 documents tailored to the MEDIQA-EVAL task. These documents served as an external evidence base for grounding subsequent model predictions.

Both proprietary and open-source multimodal systems were evaluated under identical experimental conditions, including GPT-5-mini, MedGemma-27B, MedGemma-1.5-4B, Qwen3-VL-32B, and Qwen3-VL-30B-Thinking. Each model was tested with and without retrieval-augmented generation (RAG) to quantify the impact of evidence grounding on evaluation performance.

In addition, the team implemented a local multimodal alternative that integrates DermLIP visual encoders with SBERT and KeyBERT for semantic alignment, followed by Qwen-based reasoning. This architecture enables clinically grounded scoring through joint visual–textual processing in a fully local pipeline.

6.3. MedAware Team

The MedAware team formulated medical response evaluation as a structured JSON prediction task

using Qwen3-VL vision–language models. All evaluation metrics were jointly predicted in a single decoding pass, conditioned on the clinical case, candidate response, gold references, and associated images. Training Procedure:
(i) *Supervised Fine-Tuning (SFT).* Qwen3-VL (8B and 32B) was fine-tuned using LoRA (rank 16, $\alpha = 32$), optimizing cross-entropy over output tokens. Model selection used 5-fold cross-validation stratified by encounter ID, evaluated via Kendall, Pearson, and Spearman correlations. The 32B LoRA model consistently outperformed smaller and fully fine-tuned variants on English metrics.
(ii) *Group Relative Policy Optimization (GRPO).* Following DeepSeek-R1, GRPO was applied on top of the SFT model. For each instance, eight outputs were sampled. The reward combined (i) exact-match accuracy with human ratings and (ii) a small formatting bonus for structured reasoning using `<think>...</think>`. GRPO optimized group-relative advantages without a separate value function.

This approach enables joint multi-metric prediction with structured outputs, combining parameter-efficient fine-tuning and reinforcement learning to better align predictions with human judgments.

6.4. BDI Team

The BDI team proposed a system-level approach that departed from the dominant fine-tuning paradigm commonly adopted in free-text medical evaluation. Instead of training or adapting task-specific language models, the team implemented an agent-based evaluation framework that orchestrated multiple existing components via API access to an off-the-shelf language model.

The system was built around a ReAct-style agent that performed structured reasoning guided by task-specific prompts. The agent integrated multimodal retrieval of similar clinical encounters, auxiliary machine learning–based scoring signals, AI-generated VQA references, and optional image augmentation tools, without any additional pre-training or fine-tuning. This design investigated the feasibility of modular, compositional agents as a lightweight and computationally efficient alternative for free-text medical QA evaluation, reflecting broader trends toward agentic systems in applied NLP and multimodal reasoning.

6.5. hgkai26 Team

The hgkai26 team implemented a multimodal LLM-as-a-Judge evaluation framework built on GPT-4o within the OpenAI Evals infrastructure to systematically assess system-generated responses for dermatology and wound care queries. Rather than relying on manual annotation or task-specific classifier training, the framework leveraged a rubric-guided large language model to perform structured, criteria-based grading of candidate answers.

Each evaluation instance consisted of a user query, an associated clinical image, and a candidate response. For every query, three candidate responses were independently evaluated using standardized prompts that encoded predefined scoring rubrics. The judge model assigned dimension-level scores across key quality attributes—including completeness, factual accuracy, relevance, and writing quality—using a discrete scoring scale of 0, 0.5, 1. In addition to these granular assessments, the system produced an aggregate overall score and a binary disagreement indicator to capture potential inconsistencies in evaluation outcomes. The evaluation pipeline was fully automated, generating structured quantitative outputs that enabled consistent, reproducible comparison of candidate responses. This design facilitated scalable benchmarking of multimodal medical question-answering systems by operationalizing helpfulness and response quality through rubric-based, model-driven assessment.

7. Conclusion

In this paper, we introduced the MEDIQA-EVAL 2026 shared task on evaluation metrics for medical multimodal question answering, motivated by the growing gap between rapid advances in multimodal clinical QA systems and the limited reliability of existing evaluation methods. Prior work has shown that existing automatic metrics, whether reference-based or LLM-based, do not consistently align with clinician judgments, particularly in specialized medical settings. This gap is especially evident in open-ended MMQA, where quality depends on factual accuracy, visual relevance, completeness, and overall coherence rather than surface similarity alone. By shifting the focus to evaluating the evaluators, our shared task establishes a clinician-grounded benchmark for automatic metric assessment in medical multimodal QA.

The participating systems span diverse methodological directions, including fine-tuned vision-language models, reinforcement learning, retrieval-augmented judging, metric-specific classifiers, agent-based pipelines, and prompting-based LLM-as-a-Judge frameworks. Together, the results provide a comparative overview of current approaches to multilingual and multimodal clinical evaluation. Through clinician-rated data and multidimensional quality criteria, this shared task offers a foundation for developing more reliable and clinically meaningful evaluation methods. Future work will focus on expanding the datasets by increasing the number of responses and ratings for English and Chinese.

Appendix: Prompt Templates

MEDIQA LLM-as-a-Judge: English Prompt

```
You are given a patient query, relevant medical images, a list of reference responses, and a candidate response. The reference responses were written by practicing medical doctors.
Please evaluate the candidate response according to the following metrics, using the same criteria as a practicing medical doctor.
Metrics: - disagree_flag (0 or 1): 1 if the candidate response clearly disagrees with the expert reference responses on key medical facts; 0 otherwise. - completeness (0.00 to 1.00): How completely the response answers the patient's question. - factual-accuracy (0.00 to 1.00): Whether the medical information is factually accurate. - relevance (0.00 to 1.00): Whether the response is relevant to the patient's question and images. - writing-style (0.00 to 1.00): Clarity, professionalism, and appropriateness of tone.
- overall (0.00 to 1.00): Overall quality considering completeness and factual accuracy primarily.
Calibration guidance: - Responses that closely match most reference responses in diagnosis and management should receive scores ≥ 0.85 for completeness, factual-accuracy, and overall. - Responses that partially match references should receive scores around 0.40-0.70.
- Responses that contradict references or omit key information should receive scores ≤ 0.30.
Constraints: - Score each metric independently. - A high writing-style score must not increase factual-accuracy or completeness.
- overall must not be higher than both completeness and factual-accuracy. - If either completeness or factual-accuracy ≤ 0.50, overall should also be ≤ 0.50.
Use the reference responses and the medical images as the primary evidence for your evaluation.
Output requirements: - For disagree_flag: output 0 or 1 only. - For all other metrics: output a continuous score in [0.00, 1.00] with two decimals.
Return ONLY a valid JSON object with exactly the following keys: disagree_flag, completeness, factual-accuracy, relevance, writing-style, overall All values must be numeric.
```

MEDIQA LLM-as-a-Judge: Chinese Prompt

```
你将看到一个患者问题、相关医学图像、若干参考回答，以及一个候选回答。参考回答由医学专家提供。
请按照中国医学院受训的医学专家的评估标准，对候选回答进行评分。
评分指标：- factual-consistency-wgold (0.00 到 1.00)：表示候选回答在医学事实上与金标准参考回答的一致程度。1.00 表示在诊断、判断或处理建议上与参考回答完全一致；0.00 表示与参考回答存在明显矛盾或医学错误；中间值表示部分一致或遗漏部分要点。
- writing-style (0.00 到 1.00)：表示回答在中文语言表达上的专业性、清晰度和得体程度。1.00 表示语言自然、专业，符合中文医学表达习惯；0.00 表示表达不清晰、不专业或明显不合适；中间值表示基本通顺但略显生硬或不够规范。
重要说明：- factual-consistency-wgold 只评估医学事实是否与参考回答一致，不要考虑语言是否流畅。- writing-style 只评估语言表达质量，不要考虑医学内容是否正确。- 两个评分必须相互独立，不要相互影响。
评分参考：- 若候选回答在关键医学结论上与参考回答高度一致，factual-consistency-wgold 通常应 ≥ 0.85。- 若仅部分一致或遗漏部分要点，通常在 0.40-0.70 之间。- 若存在关键矛盾或明显错误，通常 ≤ 0.30。
医学图像应作为判断医学事实一致性的重要依据。
输出要求：- 所有评分必须是 [0.00, 1.00] 范围内的连续数值（保留两位小数）。
请仅返回一个 JSON 对象，且必须包含以下键：factual-consistency-wgold, writing-style 所有值必须是数值。
```


Acknowledgements

We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP organizers for their support for the shared task. We also thank the ClinicalNLP reviewers, our annotation team, and the participating teams who contributed to the success of the shared task through their interesting approaches and strong engagement.

Limitations

This shared task does not cover all possible methods for multimodal question answering evaluation. In addition, the test set is limited in both size and domain coverage, focusing primarily on dermatology. As a result, the findings may not generalize to other medical specialties or broader clinical settings. Further validation on larger and more diverse datasets is necessary to assess the robustness and generalizability of the best-performing approaches.

8. Bibliographical References

Asma Ben Abacha, Vivek V. Datla, Sadid A. Hasan, Dina Demner-Fushman, and Henning Müller. 2020. Overview of the vqa-med task at imageclef 2020: Visual question answering and generation in the medical domain. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 1081–1088. CEUR-WS.org.

Asma Ben Abacha, Wen-wai Yim, George Michalopoulos, and Thomas Lin. 2023. An investigation of evaluation methods in automatic medical note generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pages 2575–2588. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6556–6576. Association for Computational Linguistics.

Haritha Gangavarapu. 2026. hgkai26 at mediqa-eval 2026: Automatically evaluating the accuracy of ai-generated responses in visually driven specialties. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Ziqi Hao and Pengbo Liu. 2026. Medaware at mediqa-eval 2026: Vision-language model fine-tuning with logprob-based score calibration for medical response evaluation. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *CoRR*, abs/2003.10286.

Primoz Kocbek, Valentina Carbonari, Pierangelo Veltri, Pietro Hiram Guzzi, and Gregor Stiglic. 2026. Slocal-net at mediqa-eval 2026: Investigating the impact of reasoning and external context on medical answer grading. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

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.

Xinzhe Peng, Liyuan E, Kun Feng, Jielin Li, Yuxuan Tang, and Zhao Li. 2026. Suat-bmi at mediqa-eval 2026: An ensemble approach to language models as judges for automatic rating of medical responses. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Annual Meeting of the Association for Computational Linguistics*.

Justin Xu, Zizheng Zhang, Augustine Luk, Benjamin Khong, Haochen Cui, Samuel Hwang, Alyssa Pradhan, Kevin Yuan, and David W. Eyre. 2026. Bdi at mediqa-eval 2026: A react-style multimodal agent for fine-grained medical response

assessment. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026*, Palma, Mallorca, Spain. European Language Resources Association.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2026. MORQA: benchmarking evaluation metrics for medical open-ended question answering. In *LREC, Mallorca, Spain, 13-15 May 2026*.

Wen-wai Yim, Asma Ben Abacha, Noel Codella, Roberto A. Novoa, and Josep Malvehy. 2025a. Overview of the MEDIQA-MAGIC task at imageclef 2025: Multimodal and generative telemedicine in dermatology. In *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025*, volume 4038 of *CEUR Workshop Proceedings*, pages 2233–2240. CEUR-WS.org.

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. 2025b. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *J. Biomed. Informatics*, 170:104888.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024a. Overview of the MEDIQA-M3G 2024 shared task on multilingual multimodal medical answer generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop, ClinicalNLP@NAACL 2024, Mexico City, Mexico, June 21, 2024*, pages 581–589. Association for Computational Linguistics.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, and Fei Xia. 2024b. Overview of the MEDIQA-MAGIC task at imageclef 2024: Multimodal and generative telemedicine in dermatology. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 1456–1462. CEUR-WS.org.

Wen-wai Yim, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2025c. Overview of the MEDIQA-WV 2025 shared task on woundcare visual question answering. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pages 17–21, Virtual. Association for Computational Linguistics.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024c. Dermavqa: A multilingual visual question answering dataset for dermatology. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part V*, volume 15005 of *Lecture Notes in Computer Science*, pages 209–219. Springer.

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.