BDI at MEDIQA-EVAL 2026: A ReAct-Style Multimodal Agent for Fine-Grained Medical Response Assessment

Justin Xu[1], Zizheng Zhang[1], Augustine Luk[1], Benjamin Khong[2]
Haochen Cui[1], Samuel Hwang[3], Alyssa Pradhan[1], Kevin Yuan[1], David W Eyre[1]
[1]University of Oxford, [2]University of California, Los Angeles, [3]University of California, Berkeley

Abstract

Free-text evaluation of multimodal clinical question answering (QA) systems remains a central challenge in medical NLP due to the complexity of medical knowledge, the necessity of integrating visual and textual information, and the limitations of existing automatic evaluation metrics for open-ended outputs. In this work, we present a training-free, agentic evaluation framework that formulates response scoring as evidence-guided orchestration of components rather than a task requiring conventional end-to-end fine-tuning of underlying LLMs/VLMs. Our ReAct-style evaluator combines (i) structured reasoning, (ii) multimodal retrieval of similar encounters, (iii) auxiliary explainable feature-based regression models that provide numeric priors and human-interpretable signals, (iv) VLM-generated visual QA references for comparison, and (v) optional image augmentation tools. Unlike standard LLM-as-a-judge approaches that rely on direct generative scoring, our agent decomposes evaluation into modular stages of evidence acquisition, structured feature modeling, and integrative reasoning. We apply this architecture to the MEDIQA-EVAL shared task – a multimodal, multilingual clinical evaluation challenge that assesses system-generated answers for patient queries paired with images along multiple clinical quality dimensions. We report results across both English and Chinese tracks, comparing against baseline prompting methods, and discuss the feasibility and limitations of lightweight agentic systems for clinical QA evaluation.

Keywords: evaluation, clinical applications, multimodal, multilingual, agentic AI

1. Background & Introduction

Digital messaging platforms have become an increasingly important component of modern healthcare delivery, enabling remote patient-provider communication and improving access to care while reducing logistical and financial barriers (Shickel et al., 2018; Topol, 2019). In visually driven medical specialties such as dermatology and wound care, patient queries frequently include images alongside free-text descriptions of symptoms, disease progression, or prior treatments (Yim et al., 2024c, 2025b). Automatically generating draft responses to such multimodal queries has thus become a promising approach for alleviating clinician workload and supporting scalable telemedicine.

Recent advances in large language models (LLMs) and vision-language models (VLMs) have substantially improved the fluency and expressiveness of generated medical text, as well as the ability to reason over visual inputs (Li et al., 2023; Singhal et al., 2023). However, reliably *evaluating* the quality of these free-text, multimodal responses remains an open research challenge. Unlike closed-form question answering (QA) or classification tasks, medical free-text responses may admit multiple valid answers reflecting different clinical opinions, levels of detail, or stylistic preferences. As a result, traditional automatic evaluation metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) often correlate poorly with expert judgment in clinical settings (Peng et al., 2017; Yim et al., 2025a). This problem is further compounded in multimodal scenarios, where the correctness of a response depends on appropriate interpretation of clinical images in conjunction with textual context.

A growing body of work has emphasized the need for specialized evaluation frameworks that better align with human clinical assessments. Prior efforts have explored learned evaluation models, rubric-based scoring, and LLM-as-a-judge paradigms to assess factuality, completeness, and clinical relevance of generated outputs (Liu et al., 2023; Fu et al., 2023; Xu et al., 2025a). While LLM-as-a-judge approaches have demonstrated strong empirical performance across many text generation tasks, they typically rely on direct generative scoring in a single prompt-driven inference step. Such setups offer limited transparency into the reasoning process underlying a score and provide little structured evidence for downstream auditing or calibration. Moreover, scores may also be overly influenced by surface-level properties such as verbosity or stylistic alignment, and judgments may vary with prompt phrasing or contextual framing rather than underlying clinical correctness.

In parallel, several shared tasks have advanced multimodal medical QA by focusing on answer generation in dermatology and wound care, including DermaVQA and WoundcareVQA (Yim et al., 2024c, 2025b). While these benchmarks have driven progress in multimodal reasoning, previous tasks have primarily focused on generation perfor-

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 141–152
16 May 2026.

mance and provide limited insight into the broader problem of response evaluation.

The MEDIQA-EVAL shared task (Ben Abacha and Yim, 2026), organized as part of the LREC 2026 ClinicalNLP Workshop, addresses this gap by shifting the focus from answer generation to evaluation of system-generated free-text responses. In this task, participants are given patient questions paired with one or more clinical images and are asked to assign quality scores to candidate answers produced by external systems. The task is both multimodal and multilingual, encompassing English and Chinese data, and evaluates responses along multiple clinically motivated dimensions, including overall quality, factual accuracy, completeness, relevance, and writing style. By grounding evaluation in expert human ratings, MEDIQA-EVAL provides a rigorous testbed for developing and analyzing automated evaluation methodologies in realistic settings.

In this paper, we describe Team BDI's submission to the MEDIQA-EVAL shared task. Our approach departs from the dominant paradigm of fine-tuning task-specific models for evaluation. Instead, we propose a modular, agent-based framework that treats evaluation as evidence-guided orchestration. The system is built around a ReAct-style agent (Yao et al., 2023) using an out-of-the-box LLM that integrates structured reasoning guided by task-specific prompts, multimodal retrieval of similar clinical encounters, auxiliary machine learning-based scoring signals, VLM generated references, and optional image augmentation tools. This design reflects recent trends toward agentic and compositional systems in applied natural language processing and explores their viability as lightweight and flexible alternatives for multimodal clinical QA evaluation.

Our contributions are three-fold:

- We introduce a ReAct-style, training-free evaluation framework for multimodal clinical QA that decomposes scoring into explicit stages of structured rubric-based reasoning, evidence retrieval, and integrative decision-making, rather than relying on monolithic end-to-end grading.
- We design an interpretable scoring mechanism that augments LLM-based reasoning with structured features and regression-based numeric priors, enabling more transparent, auditable, and potentially calibratable judgments compared to single-step LLM-as-a-judge approaches.
- We provide an empirical analysis on the multilingual MEDIQA-EVAL benchmark, showing that inference-time modular orchestration offers a viable alternative for multimodal clinical QA evaluation across English and Chinese settings that does not require any additional pre-training or fine-tuning of underlying foundation models.

2. Related Works

2.1. Medical Question Answering

Medical QA has evolved from traditional text-only retrieval or multiple-choice formats to increasingly complex multimodal tasks that integrate both visual and textual information (Lin et al., 2023). Early biomedical QA benchmarks focused on answering factual questions over text corpora or structured clinical data (Ben Abacha et al., 2018; Nentidis et al., 2024). However, recent work has expanded QA to incorporate multimodal reasoning, where systems must jointly interpret clinical images and accompanying questions to produce accurate answers (Demirhan and Zadrozny, 2024; Zhang et al., 2025). For example, large-scale medical visual QA efforts have targeted generation of natural language responses grounded in radiological, pathological, or endoscopic images paired with clinical queries (Zhang et al., 2024; Gautam et al., 2025).

2.2. Free-Text Generation and Evaluation

Free-text generation in medical QA introduces unique evaluation challenges due to variability in valid clinical responses and nuanced medical content. Several shared tasks have focused on multimodal free-text generation conditioned on clinical images and questions (Yim et al., 2024b; Xu et al., 2024). Similarly, the MEDIQA-WV and MEDIQA-MAGIC tasks have emphasized wound care and dermoscopy QA, requiring systems to produce fluent, clinically appropriate text accompanied by structured output where applicable (Yim et al., 2024a, 2025c).

Despite progress in generation, evaluation remains an active area of research. Generic automatic metrics often correlate poorly with clinical relevance and accuracy, leading to work on specialized evaluators and metrics (Van Veen et al., 2024; Ostmeier et al., 2024; Zhao et al., 2024; Xu et al., 2025b; Delbrouck et al., 2025). This underscores the need for evaluation benchmarks like MEDIQA-EVAL that directly target free-text quality in clinical QA.

2.3. Dermatology and Wound Care

Recent work has also explored foundation models and specialized datasets tailored to dermatology and wound care domains. In dermatology, the MM-Skin dataset enhances VLM capability by deriving large-scale image-text pairs from textbooks and

clinical sources, enabling richer skin disease representation and QA performance (Zeng et al., 2025). Other dermatology visual QA efforts such as DermaVQA provide multilingual benchmarks for skin condition reasoning (Yim et al., 2024c). In wound care, the WoundcareVQA dataset supports multimodal QA with structured wound attributes and clinical queries (Yim et al., 2025b), and recent retrieval augmented generation methods have been shown to be effective in the MEDIQA-WV shared task framework (Karim and Uzuner, 2025).

3. MEDIQA-EVAL Shared Task

The MEDIQA-EVAL shared task evaluates automatic methods for scoring free-text, multimodal clinical QA responses. Given a patient question paired with one or more clinical images and a set of candidate system-generated answers, participating systems assign numerical quality scores intended to approximate expert human judgements. The task is multilingual (English and Chinese) and focuses on visually driven clinical domains, specifically dermatology and wound care.

The benchmark builds on two previously released multimodal datasets: *DermaVQA* (referred to as `iiyi`), a dermatology visual question answering dataset, and *WoundcareVQA*, which focuses on wound care encounters. Both datasets consist of online patient queries accompanied by clinical images and clinician-authored responses. For the shared task, each encounter is paired with three candidate system outputs to be evaluated. A summary of the dataset composition used in the challenge is shown in Table 1.

Dataset	Split	#Encounters	#Systems	#Human
WoundcareVQA	Dev	105	315	210
	Test	93	279	279
DermaVQA (iiyi)	Dev	56	168	417
	Test	100	300	926

Table 1: Summary of datasets used in MEDIQA-EVAL. Each encounter is associated with three system-generated candidate responses and one or more human responses. Both English (EN) and Chinese (ZH) versions are provided for both development and testing splits.

Human expert annotations serve as the reference standard for evaluation. For English data, candidate responses are rated along multiple clinically motivated axes:

- **Completeness**: extent to which the response adequately addresses all clinically relevant aspects of the query.
- **Factual Accuracy**: clinical correctness of the medical information provided.
- **Relevance**: degree of alignment between the response and the patient's question.
- **Writing Style**: clarity, coherence, and appropriateness of the language used.
- **Overall Quality**: holistic assessment of the response's clinical usefulness and reliability.

Each of these five metrics is discretized to $\{0, 0.5, 1\}$, corresponding to poor, partial, or strong performance. In addition, the English data includes a binary **disagreement flag** indicating whether the evaluator *disagrees* with the overall clinical judgment or message conveyed in the response. Chinese samples are evaluated using a reduced set of criteria focusing on only factual consistency with human responses and writing quality. These dimensions are likewise discretized to $\{0, 0.5, 1\}$ and do not include a disagreement flag.

The development data was provided with multiple expert judgements per candidate. However, each candidate in the test data was independently annotated by two experts, with a conflict resolution process applied by the task organizers for cases of disagreement to produce the final reference scores used for evaluation.

Submitted automatic scoring systems are evaluated by measuring their correlation with human ratings. For each metric, Kendall's τ, Pearson's r, and Spearman's ρ are computed between system-generated scores and human judgements. These correlation coefficients are averaged to produce a metric-level score. Final leaderboard rankings are determined by averaging metric-level scores across all metrics within each language, yielding separate aggregate scores for English and Chinese. This evaluation protocol emphasizes relative ranking and calibration of candidate responses rather than absolute score prediction, encouraging systems to capture clinically meaningful distinctions among candidate answers.

4. Methods

4.1. Task Formulation

We formulate MEDIQA-EVAL as a multimodal response scoring problem. Let an encounter be defined as:

$$e = \langle q, I, M \rangle$$

where q denotes a patient query expressed as free text, $I = \{i_1, \ldots, i_{|I|}\}$ is a set of associated clinical images, and M represents optional structured metadata describing the encounter (*e.g.*, anatomical location or wound type). For each encounter e, a fixed set of candidate system responses $C = \{c_1, c_2, \ldots, c_K\}$ is provided, where

each c_k is a free-text answer generated by an external system.

The objective of the task is to assign numerical quality scores to each candidate c_k along a predefined set of evaluation dimensions. Let $\mathcal{D}$ denote the set of evaluation metrics, which differs by language (*i.e.*, six metrics for English and two for Chinese). The desired output of an automatic evaluation system is a function:

$$f : (e, c_k, d) \rightarrow \hat{y}_{k,d}$$

where $\hat{y}_{k,d} \in \mathbb{R}$ is the predicted score for candidate c_k on metric $d \in \mathcal{D}$. These predictions are evaluated by their rank and linear correlation with expert human judgements rather than absolute error, emphasizing relative ordering and calibration across candidates.

4.2. Evaluation Agent Overview

To address this task, we implement an *agent-based evaluation framework* that treats response scoring as a structured reasoning and orchestration problem rather than a purely parametric prediction task. At a high level, the agent operates as a policy:

$$\pi_\theta : (e, C) \rightarrow \{\hat{Y}_1, \ldots, \hat{Y}_K\}$$

where $\hat{Y}_k = \{\hat{y}_{k,d} \mid d \in \mathcal{D}\}$ denotes the vector of predicted metric scores for candidate c_k. The policy π_θ is instantiated by `gpt-5-2025-08-07` out-of-the-box, conditioned on a system prompt and augmented through external tools (Figure 1).

Following the ReAct paradigm, the agent interleaves natural language reasoning with tool invocations to gather auxiliary evidence before producing final scores. Rather than embedding all task-specific knowledge in model parameters, the agent dynamically queries specialized modules that expose structured signals relevant to clinical evaluation. This design emphasizes modularity and interpretability and avoids additional model fine-tuning.

Concretely, for each encounter e, the agent performs an iterative reasoning process: $s_{t+1} = \text{LLM}(s_t, o_t)$, where s_t denotes the agent's internal reasoning state at step t, and o_t represents observations returned by external tools. The process terminates when the agent emits a complete set of metric predictions for all candidates in C.

4.3. Model Context Protocol Tools

The agent interacts with external components through the Model Context Protocol (MCP) (Hou et al., 2025), which provides a standardized interface for invoking tools implemented as independent services. Each tool exposes a well-defined input-output contract and can be queried conditionally based on the agent's reasoning state. In our system, we employ several MCP servers corresponding to complementary sources of evidence: feature-based encounter retrieval, auxiliary machine learning-based scoring, VLM generated references, and image augmentation skills.

4.3.1. Feature-based Encounter Retrieval

Our retrieval subsystem returns a small set of previously observed encounters:

$$R(e) = \{e'_1, \ldots, e'_N\}$$

that are most relevant to an input encounter $e = \langle q, I, M \rangle$. Retrieval is implemented as a two-stage pipeline: (i) *metadata-based filtering* to produce a reduced candidate pool, followed by (ii) *embedding-based ranking* over text and images, with final fusion via reciprocal rank fusion (RRF) (Cormack et al., 2009).

Metadata Features. To enrich the datasets' sparse structured metadata, we first generated additional metadata annotations using `gpt-5.2-2025-12-11`. Specifically, for each encounter, we prompted the LLM to infer structured attributes from the query text and images, such as:

- **`query_type`**: coarse intent of the question (*e.g.*, diagnosis, management, urgency).
- **`anatomic_locations`**: normalized anatomical site(s) mentioned (list-valued).
- **`suspected_diagnosis`**: short phrase describing the likely diagnosis (scalar).
- **`contains_identifiable_feature`**: binary indicator whether the image shows an identifiable person/feature.

For metadata retrieval matching, we score store entries by counting exact matches for scalar fields and Jaccard similarity for list fields. Let M_{in} be the input metadata and M_{s} a store entry's metadata. We define the scalar-match component:

$$S_{\text{scalar}}(M_{\text{in}}, M_{\text{s}}) = \sum_{k \in \mathcal{K}_{\text{scalar}}} \mathbf{1}\big[M_{\text{in}}(k) = M_{\text{s}}(k)\big]$$

and the list-field component:

$$S_{\text{list}}(M_{\text{in}}, M_{\text{s}}) = \sum_{k \in \mathcal{K}_{\text{list}}} \frac{|A_k \cap B_k|}{|A_k \cup B_k|}$$

where A_k and B_k are the normalized sets for field k in the input and store entry respectively. The overall metadata score is the sum:

$$S_{\text{meta}}(M_{\text{in}}, M_{\text{s}}) = S_{\text{scalar}}(M_{\text{in}}, M_{\text{s}}) + S_{\text{list}}(M_{\text{in}}, M_{\text{s}})$$

We select the top $N_{\text{meta}} = 10$ store entries ranked by S_{meta} (ties broken deterministically) to form the candidate pool for the next stage.

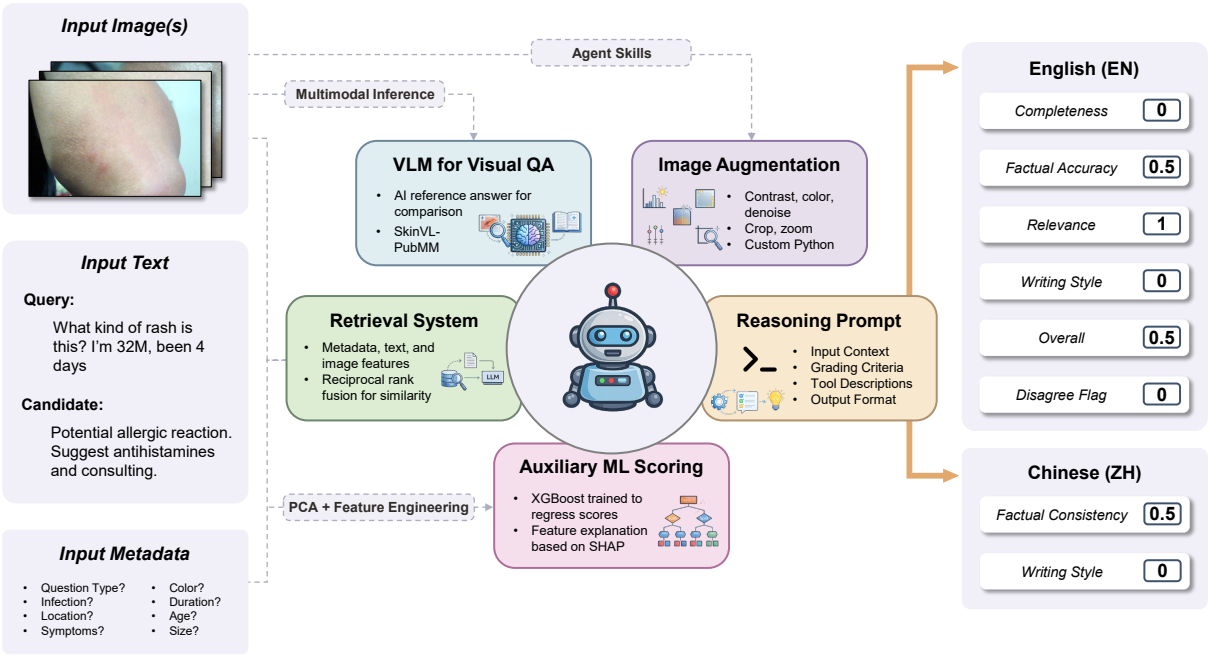

Figure 1: Schematic of the evaluation agent showing various tool modules. A representative English encounter is shown. For Chinese encounters, the input query and candidate would be translated, and the output would only consist of two metrics.

Textual Features. For textual similarity, we embed full queries and candidate answer text using `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2019). In addition to dense embeddings, we perform light clinical entity extraction using *regex* and keyword lists (*e.g.*, age, duration, and anatomical terms) to provide interpretable overlap features. Given a query embedding $\mathbf{q} \in \mathbb{R}^d$ and a store text embedding $\mathbf{t} \in \mathbb{R}^d$, we compute cosine similarity:

$$\text{sim}_{\text{text}}(\mathbf{q}, \mathbf{t}) = \frac{\mathbf{q} \cdot \mathbf{t}}{\|\mathbf{q}\|\|\mathbf{t}\|}$$

Image Features. For visual similarity we include two complementary signals:

- *MedImageInsight* embeddings (Codella et al., 2024) extracted via an internal endpoint. These capture clinically relevant visual representations and serve as the primary medical image embedding source.
- *ViT CLIP* embeddings (Radford et al., 2021) to provide a general-purpose visual-text alignment signal.

We additionally compute descriptive image statistics (*e.g.*, RGB mean/std and HSV means) and texture descriptors (*e.g.*, patch-level variance and simple saliency patch counts). For multi-image encounters we represent the visual evidence as the set of per-image embeddings $\{\mathbf{i}_1, \ldots, \mathbf{i}_m\}$. The image-to-image similarity between input images and a store encounter is computed by taking the maximum pairwise cosine similarity across all image pairs:

$$\text{sim}_{\text{image}}(I, I_s) = \max_{u \in I} \max_{v \in I_s} \frac{\mathbf{i}_u \cdot \mathbf{j}_v}{\|\mathbf{i}_u\|\|\mathbf{j}_v\|},$$

which naturally handles differing image counts per encounter by focusing on the best-matching pair.

Ranking and Fusion. After the metadata filtering stage produces candidate encounters $\mathcal{C}_{\text{meta}}$, we produce two ranked lists over $\mathcal{C}_{\text{meta}}$: a text-based ranking using sim_{text} and an image-based ranking using $\text{sim}_{\text{image}}$. Each ranking is an ordered list with ranks $r_{\text{text}}(e')$ and $r_{\text{image}}(e')$ for $e' \in \mathcal{C}_{\text{meta}}$. We fuse the two rankings using RRF:

$$\text{score}_{\text{RRF}}(e') = \sum_{m \in \{\text{text,image}\}} \frac{1}{k + r_m(e')},$$

where k is an RRF parameter (we use $k = 60$ in our implementation). The fused ordering is obtained by sorting $\text{score}_{\text{RRF}}$ in descending order, and the top N fused encounters are returned to the agent as exemplars $R(e)$. In practice, we return the top 3 fused exemplars.

4.3.2. Auxiliary Machine Learning-based Scoring

To supply the agent with compact, explainable numeric reference signals, we trained auxiliary

gradient-boosted decision tree models (XGBoost (Chen and Guestrin, 2016)) that predict metric-specific scores. These models operate on a dense, precomputed feature representation $\phi(e, c_k) \in \mathbb{R}^M$ constructed from the text, image, and metadata pipelines described above. For each evaluation metric d (*e.g.*, `completeness`, `factual-accuracy`, and `overall`), we fit a regression model:

$$g_d : \mathbb{R}^M \to \mathbb{R}, \qquad \tilde{y}_{k,d} = g_d\big(\phi(e, c_k)\big)$$

where $\tilde{y}_{k,d}$ is the predicted score used as an auxiliary signal by the agent.

Feature Engineering. The feature matrix is created in two stages. First, we assemble a base row for every (encounter, candidate) pair containing:

- **Text features:** query and candidate dense embeddings reduced with PCA to $d_{\text{PCA}} = 32$ dimensions, discrete query type, entity counts, candidate length, and punctuation statistics.
- **Image features:** mean-pooled per-image *MedImageInsight* embeddings reduced with PCA to $d_{\text{PCA}} = 32$, saliency/patch statistics, and image tag one-hots.
- **Metadata features:** one-hot / count encodings for anatomic locations, wound type, and author/rater statistics.

Second, we augment the base matrix with interpretable and interaction features, such as Flesch reading ease (English), average tokens-per-sentence, counts of medical terms, action recommendation, harmful-recommendation heuristics, cross-modal cosine similarities (between query-image), mean/max textual cosine similarities (between candidate-gold), and triangle-style features (cosine similarity of `query_candidate` minus cosine similarity of `image_candidate`).

Training Objective and Explainability. For each metric d on the validation set, we use the mean human rating (aggregated across raters for that candidate) as the regression target. Models were trained using XGBoost with squared error objective. We train separate models per metric and validate using held-out validation folds derived from the provided development split.

For each prediction $\tilde{y}_{k,d}$, we also compute SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) values to extract the top-K explainable features (excluding PCA/embedding dimensions) that contributed most to the model output.

Agentic Model Usage. During per-encounter evaluation, the ReAct agent invokes the XGBoost MCP to retrieve $\{\tilde{y}_{k,d}\}_d$ and the associated top-5 SHAP features for each candidate in natural language form to guide downstream reasoning, which was previously shown to aid in feature engineering for tabular clinical prediction tasks (Zhang et al., 2026). The agent conditions on these signals; however, it is free to accept, down-weight, or contradict the auxiliary predictions. We designed this interaction to provide a structured, interpretable corrective signal (in the spirit of explainability-driven model steering) without hard-wiring the final outputs to the auxiliary regressor.

4.3.3. VLM-based Visual QA Reference

We also incorporate a VLM-based visual QA reference. Given an encounter $e = \langle q, I, M \rangle$, we generate a VLM-based answer using a pre-trained multimodal foundation model specialized for dermatological reasoning (`zwq803/SkinVL-PubMM` (Zeng et al., 2025)). The model takes the query text and associated clinical images as input and produces a free-text response intended to approximate a clinically appropriate answer.

The generated response is not treated as ground truth. Instead, it serves as an additional reference signal that the evaluation agent can compare against candidate responses. The agent may assess semantic similarity between a candidate c_k and the generated answer, or use discrepancies between them as evidence when evaluating factual accuracy or completeness.

4.3.4. Image Augmentation Skills

Agent skills (Anthropic, 2025) expose a lightweight image augmentation toolkit that the agent may invoke when visual evidence is insufficient (*e.g.*, low contrast, poor lighting, or small lesion sizes) or when a localized patch requires closer inspection. Augmentations are treated as deterministic image-to-image transformations $i'_j = T(i_j; \psi)$, where i_j is an input image, T is an augmentation operator selected by the agent, and ψ denotes operator parameters (*e.g.*, Contrast Limited Adaptive Histogram Equalization (CLAHE (Pizer et al., 1990)) clip limit or gamma). Augmented images are returned to the agent for inspection and reasoning.

Invocation and Augmentations. The agent supplies a short, self-contained Python code snippet that reads the input image, applies image-processing operations (using `Pillow`, `cv2`, and/or `numpy`), and writes the result. The snippet runs inside a sandboxed subprocess with no external network or arbitrary filesystem access. The agent then loads the augmented image and continues

reasoning. This design keeps augmentations explicit, auditable, and reversible. We implemented and recommended a small set of clinically useful operations that the agent may combine sequentially:

- **Contrast enhancement / CLAHE.** Apply CLAHE to local image tiles to improve visibility of subtle texture or color differences.
- **Edge / border emphasis.** Sobel or Laplacian filters (optionally blended with the original image) to highlight lesion margins.
- **Color normalization / channel emphasis.** Scale or reweight RGB/Hue channels to emphasize erythema or pigmentation (*e.g.*, amplify red channel or threshold HSV ranges).
- **Denoising.** Apply bilateral or non-local means denoising prior to contrast steps to avoid amplifying noise.
- **Crop / zoom to region-of-interest (ROI).** Crop and resize to create a close-up patch for finer inspection.

The agent decides to invoke augmentation based on internal signals and metadata: *e.g.*, the agent inspects `contains_identifiable_feature` or computed image statistics (low contrast, low mean brightness, or low saliency counts) and may request an augmentation if these indicators suggest improvement is possible. The agent is instructed to keep augmentations *moderate* to avoid introducing artifacts that could mislead clinical interpretation. In practice, these augmentations modestly improve the agent's ability to (i) detect lesion borders and exudate, (ii) disambiguate pigmentation vs. shadowing, and (iii) increase the signal-to-noise ratio of patch-level texture statistics. Because augmentations are fast (simple `OpenCV`/`Pillow` operations) and executed only on demand, they provide a computationally cheap way to increase visual interpretability without retraining visual encoders.

5. Results

5.1. English Subtask

Table 2 summarizes the English leaderboard results. Our system ranked 4th out of 5 teams on the aggregate mean score across all metrics. Only one participating team surpassed the organizer-provided baseline on the mean-of-all-metrics score. The strongest systems in this track adopted supervised fine-tuning strategies using the development data. In contrast, our submission relied entirely on prompting and modular tool orchestration without additional model training.

Although our aggregate ranking was lower, several metric-level patterns are worth highlighting. Our system performed competitively on the `writing-style` metric and achieved relatively strong results on the `overall` metric compared to other non-fine-tuned approaches. This suggests that the agent's structured reasoning and auxiliary signals were effective at capturing holistic quality and stylistic alignment.

However, performance was substantially lower on the `completeness` metric. We observed that the agent tended to assign lower completeness scores to concise human-generated answers, likely reflecting a verbosity bias inherent to LLMs, especially if the agent heavily conditioned its scores on visual QA references generated by LLMs. Because `completeness` contributes directly to the aggregated mean with equal weight to other metrics, it had a noticeable impact on the overall ranking.

To better understand the contribution of individual modules, we conducted iterative internal experiments during development. Starting from a baseline prompting setup (GPT-4o; for "All Metrics", EN: 0.186, ZH: 0.125), we incrementally added (i) retrieval, (ii) auxiliary XGBoost scoring, (iii) LLM generated references, and (iv) image augmentation tools. Each addition produced measurable improvements in correlation on our held-out set of the development split, indicating that the MCP-based modular components contribute positively even without explicit fine-tuning.

Rank	Team	All Metrics	Overall Metric
1	SUAT-BMI	0.482	0.545
2	SloCal-Net	0.416	0.466
3	MedAware	0.391	0.526
4	**BDI**	**0.369**	**0.439**
5	hgkai26	0.213	0.244
–	MEDIQA (Baseline)	0.451	0.555

Table 2: MEDIQA-EVAL leaderboard results for the English subtask.

5.2. Chinese Subtask

Table 3 presents the Chinese results. Our system ranked 2nd out of 3 teams. Unlike the English track, performance differences among teams were comparatively smaller. We observed particularly strong performance on the `factual-consistency-wgold` metric, indicating that the retrieval-augmented reasoning and candidate-gold similarity features transfer well to Chinese.

In contrast, the system underperformed on the `writing-style` metric. One plausible explanation is that stylistic expectations for Chinese online medical responses differ from the model's dominant training distribution. Since our system relies

on inference-time prompting rather than language-specific fine-tuning, stylistic calibration may be less precise in Chinese. Nevertheless, the narrow performance gap suggests that prompting-based approaches with multilingual models can remain competitive without language-specific retraining.

Overall, the Chinese results highlight a potential advantage of modular, inference-driven systems: a single architecture can generalize across languages with minimal additional engineering effort, whereas fine-tuning approaches may require separate models or additional multilingual supervision.

Rank	Team	All Metrics
1	SloCal-Net	0.294
2	**BDI**	**0.269**
3	MedAware	0.259
–	MEDIQA (Baseline)	0.277

Table 3: MEDIQA-EVAL leaderboard results for the Chinese subtask.

6. Discussion and Conclusion

We presented a modular, agent-based framework for multimodal free-text evaluation in the MEDIQA-EVAL shared task. Importantly, all components operate without additional model pre-training or fine-tuning. Task specificity is achieved through prompt design, tool orchestration, and feature engineering. This architecture enables rapid experimentation and highlights the feasibility of compositional, agentic systems for complex clinical evaluation tasks.

From a systems perspective, several design choices proved beneficial. All embeddings, metadata annotations, and document records are stored in offline indexed stores, with embedding shards and encounter-level mappings to enable efficient lookup during inference. This design minimizes redundant computation and allows scalable retrieval across encounters.

The two-stage retrieval strategy – metadata filtering followed by embedding-based ranking with RRF – reduces unnecessary cross-domain comparisons while leveraging LLM-augmented metadata to capture clinically salient attributes absent from the original dataset-supplied fields. By fusing text and image rankings, the system remains robust to modality imbalance (*e.g.*, encounters with multiple images or limited textual detail). Retrieved exemplars provide contextual human responses and historically scored candidates, supporting retrieval-augmented reasoning and structured evaluation.

The auxiliary XGBoost module complements reasoning by providing structured numeric priors and feature-level explanations via SHAP. Rather than constraining outputs, these signals act as interpretable guidance, allowing the agent to integrate model-based predictions with broader contextual reasoning. This hybrid design reflects a middle ground between fully parametric fine-tuning and purely prompt-based scoring.

Future Directions. There are several avenues for improvement. First, stronger domain-specific visual encoders, particularly dermatology- or wound-care-specialized foundation models, may yield improved image embeddings and cross-modal similarity features. Our current approach relies on general-purpose or proprietary embeddings. Domain-tuned representations could better capture subtle clinical cues.

Second, additional prompt calibration may better align LLM grading behavior with human annotation guidelines, particularly for metrics such as completeness and writing style. Explicit rubric reminders, metric-specific reasoning steps, or stylistic exemplars may improve alignment without requiring parameter updates.

Third, the evaluation problem itself could be decomposed into more constrained sub-questions. For example, instead of directly predicting a scalar completeness score, the system could assess whether key clinical components (diagnosis, management advice, risk discussion) are present and then aggregate those assessments.

Our current framework relies on a small, fixed number of retrieved exemplars and predefined fusion strategies, but we did not exhaustively explore the sensitivity of performance to these choices. Future work should conduct a rigorous ablation study examining the number of retrieved samples, metadata filtering thresholds, and text-image fusion weights, in order to better understand their contribution to final correlation scores.

Finally, synthetic data generation or expanded human annotation could help stabilize evaluation models. The development split, which the retrieval system heavily depends on, is limited in size and exhibits annotation variability. Generating additional high-quality examples, either via expert labeling or carefully controlled LLM simulation, may reduce noise sensitivity for both fine-tuning and feature-based approaches.

While our system did not achieve top leaderboard performances, the results demonstrate that modular, inference-driven agents can still be a viable method for multilingual settings with minimal retraining overhead. As multimodal evaluation benchmarks continue to evolve, compositional agent frameworks offer a flexible and extensible alternative to fine-tuned models.

7. Limitations

Dependence on noisy development data. Our retrieval-based strategy relies heavily on the quality of the development split. Although the organizers noted that the development annotations contain noise, our system retrieves at most ten metadata-matched encounters and ultimately conditions on only the top few exemplars during inference. Consequently, if the retrieved encounters contain inconsistent or noisy human ratings, this noise directly propagates into the agent's reasoning. In contrast, fine-tuning approaches implicitly learn aggregate patterns across the entire development split, potentially smoothing over local inconsistencies. Retrieval-based systems are therefore particularly sensitive to annotation quality. A possible improvement would be to curate a smaller, high-quality subset of development encounters with adjudicated ratings. However, given the inherent variability in clinician preferences, such curation may still not perfectly align with the hidden test annotations.

Inference latency and engineering overhead. The architecture depends on multiple model calls interleaved with local tool execution. Although token usage is moderate and does not require GPU-based model training, wall-clock inference time can be substantial due to sequential LLM inferences and MCP tool invocations. This may limit scalability in real-time deployment scenarios. Engineering optimizations (parallel tool execution, caching, batching, or leveraging smaller models such as `gpt-5-mini`) could substantially reduce latency. Preliminary experiments suggest that smaller frontier models can still reason effectively with tool calls, though with modest performance trade-offs.

Multilingual stylistic calibration. While the system generalizes well across languages without retraining, stylistic calibration – particularly for the Chinese writing style metric – remains imperfect. Available LLMs are predominantly trained on English-centric corpora, and stylistic expectations for online clinical communication may vary by language and cultural context. Prompt refinement or language-specific stylistic exemplars may improve alignment.

Subjectivity of clinical evaluation. More fundamentally, free-text clinical evaluation is inherently subjective. Clinician ratings depend on training background, regional practice patterns, and individual communication preferences. There is no universally standardized rubric that fully captures "correctness" or "quality" across all practitioners. As a result, even well-designed automatic systems may struggle to achieve high correlation across diverse annotator populations. This challenge extends beyond MEDIQA-EVAL and reflects a broader open research problem in evaluating generative medical AI systems.

Scope of visual representations. Our visual representations rely on general-purpose or externally trained embeddings rather than dermatology- or wound-care-specific vision models trained directly on the task distribution. Although we incorporate *MedImageInsight* and *CLIP* embeddings, further domain adaptation or use of highly specialized medical vision encoders could potentially improve cross-modal alignment.

Ultimately, while the proposed agent-based framework provides a flexible and extensible alternative to fine-tuning approaches, its performance remains sensitive to annotation quality, contextual limitations, and the intrinsic subjectivity of clinical free-text evaluation.

8. Acknowledgments

This project was funded by the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with the UK Health Security Agency (UKHSA) (NIHR207397) and supported by the NIHR Biomedical Research Centre, Oxford. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, or UKHSA. JX gratefully acknowledges joint support from Canadian Institutes of Health Research (CIHR) Project ID 202410BCB-535721-77482 (Bioinformatics and Computational Biology), Nuffield Department of Medicine (NDM), and Oxford University Press (OUP). ZZ acknowledges support from Nuffield Department of Population Health (NDPH) and OUP. AL is a recipient of the University of Oxford Croucher Scholarship. HC is supported by the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (CIFMS) Grant Number 2024-I2M-2-001-1.

9. Bibliographical References

Anthropic. 2025. Agent skills. https://platform.claude.com/docs/en/agents-and-tools/agent-skills/overview. Claude API Docs. Accessed: 2026-02-11.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2018. *Overview of the Medical Question Answering Task at TREC 2017 LiveQA*. National Institute of Standards and Technology (NIST).

Asma Ben Abacha and Wen-wai Yim. 2026. Overview of the mediqa-eval 2026 shared task on evaluation metrics in medical multimodal question answering. In *Proceedings of the 8th Clinical Natural Language Processing Workshop, ClinicalNLP@LREC 2026, Palma, Mallorca, Spain, May 16, 2026*. European Language Resource Association (ELRA).

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Noel Codella, Yu Gu, Shrey Jain, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Natieek Sangani, Sheng Zhang, Hoifung Poon, Stephanie Hyland, Shruthi Bannur, Javier Alvarez-Valle, Xue Li, John Garett, Alan McMillan, Gaurav Rajguru, Madhu Maddi, Nilesh Vijayrania, Reehan Bhimai, Nick Mecklenburg, Rupal Jain, Daniel Holstein, Naveen Gaur, Vijay Aski, Jenq-Neng Hwang, Thomas Lin, Ivan Tarapov, Matthew P. Lungren, and Mu Wei. 2024. MedImageInsight: An open-source embedding model for general domain medical imaging. *arXiv*.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 758–759. Association for Computing Machinery.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Pontes Reis, Mohamed S Muneer, Maya Varma, and Curtis Langlotz. 2025. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26813–26829. Association for Computational Linguistics.

Hilmi Demirhan and Wlodek Zadrozny. 2024. Survey of multimodal medical question answering. *BioMedInformatics*, 4(1):50–74. Publisher: Multidisciplinary Digital Publishing Institute.

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

Sushant Gautam, Michael A. Riegler, and Pål Halvorsen. 2025. Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions.

A. H. M. Rezaul Karim and Ozlem Uzuner. 2025. Multimodal retrieval-augmented generation with large language models for medical VQA.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

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

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

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.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima-López, Eulàlia Farré-Maduell, Martin Krallinger, Natalia Loukachevitch, Vera Davydova, Elena Tutubalina, and Georgios Paliouras. 2024. Overview of BioASQ 2024: The twelfth BioASQ challenge on large-scale biomedical semantic indexing and question answering.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 374–390. 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*, pages 311–318. Association for Computational Linguistics.

Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, and K.E. Muller. 1990. Contrast-limited adaptive histogram equalization: speed and effectiveness. In *[1990] Proceedings of the First Conference on Visualization in Biomedical Computing*, pages 337–345.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56. Publisher: Nature Publishing Group.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142. Publisher: Nature Publishing Group.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98.

Justin Xu, Yiming Li, Zizheng Zhang, Augustine Yui Hei Luk, Mayank Jobanputra, Samarth Oza, Ashley Murray, Meghana Reddy Kasula, Andrew Parker, and David W Eyre. 2025a. Tree-of-quote prompting improves factuality and attribution in multi-hop and medical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5605–5622, Suzhou, China. Association for Computational Linguistics.

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, David W Eyre, and Jean-Benoit Delbrouck. 2025b. RadEval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 546–557, Suzhou, China. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, and Fei Xia. 2024a. Overview of the MEDIQA-MAGIC task at ImageCLEF 2024: Multimodal and generative TelemedICine in dermatology.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025a. Morqa: Benchmarking evaluation metrics for medical open-ended question answering.

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. *Journal of Biomedical Informatics*, 170:104888.

Wen-wai Yim, Asma Ben Abacha, Yujuan Fu, Zhaoyi Sun, Fei Xia, Meliha Yetisgen, and Martin Krallinger. 2024b. 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. Association for Computational Linguistics.

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. 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 *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland.

Wenqi Zeng, Yuqi Sun, Chenxi Ma, Weimin Tan, and Bo Yan. 2025. MM-skin: Enhancing dermatology vision-language model with an image-text dataset derived from textbooks.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277. Publisher: Nature Publishing Group.

Zhilin Zhang, Jie Wang, Zhanghao Qin, Ruiqi Zhu, and Xiaoliang Gong. 2025. Efficient bilinear attention-based fusion for medical visual question answering.

Zizheng Zhang, Yiming Li, Justin Xu, Jinyu Wang, Rui Wang, Lei Song, Jiang Bian, David W Eyre, and Jingjing Fu. 2026. Medfeat: Model-aware and explainability-driven feature engineering with llms for clinical tabular prediction.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A metric for radiology report generation. Pages: 2024.06.24.24309405.