SUAT-BMI at MEDIQA-EVAL 2026: An Ensemble Approach to Language Models as Judges for Automatic Rating of Medical Responses

Xinzhe Peng*, Liyuan E*, Kun Feng, Jielin Li, Yuxuan Tang, Zhao Li
Shenzhen University of Advanced Technology
lizhao@suat-sz.edu.cn

Abstract

The MEDIQA-EVAL 2026 shared task focuses on developing automatic evaluation metrics for LLM-generated responses in dermatology and wound care. While LLMs have shown promise as judge models, the reliability of these metrics remains underexplored. In this work, we study how well judge models can approximate human expert ratings across clinical evaluation criteria. We evaluate multiple approaches, including few-shot prompting, BERT fine-tuning, and retrieval-augmented generation (RAG), and combine them in an ensemble framework. Our method achieves a correlation score of 0.481, ranking first among 41 participating teams. Our results provide insight into the reliability of LLM-based evaluation metrics and highlight their potential for scalable clinical assessment.

Keywords: Medical Question Answering, LLM, BERT, Retrieval Augmented Generation

1. Introduction

Evaluating LLM-generated responses to clinical questions remains a significant challenge due to the complexity of medical contexts and the need for nuanced semantic judgment. Such evaluations are inherently multidimensional, requiring assessment of factual accuracy, completeness, relevance, and writing quality. While expert review by trained professionals provides high-quality assessments, it is costly and difficult to scale, motivating the development of automated evaluation methods.

Recent work has increasingly explored the use of large language models (LLMs) as judge models to automatically evaluate generated outputs. These approaches rely on predefined evaluation metrics, yet it remains unclear whether such metrics, when used as criteria by LLMs, produce ratings that are consistent with human expert judgment. Understanding the reliability of these metrics is therefore a critical question.

In this work, we investigate the extent to which LLM-based judge models can approximate expert evaluations under established clinical assessment criteria. Using the MEDIQA-EVAL 2026 shared task dataset, which includes patient queries, LLM-generated responses, expert reference answers, and fine-grained ratings provided by dermatology and wound care specialists (Yim et al., 2024, 2025a,b), we frame the problem as predicting expert-assigned scores across multiple evaluation dimensions.

To study this, we propose an ensemble-based judge model that integrates multiple approaches, including supervised fine-tuning and few-shot prompting with models such as Qwen3-30B-A3B and PubMedBERT, along with retrieval-augmented generation (RAG). By selecting model configurations based on development set performance, our system aims to maximize alignment with expert ratings, achieving a final official score of 0.481.

Our results show that a carefully constructed ensemble of judge models can approximate human evaluation to a meaningful degree, while also revealing variability in how well different metrics are captured. These findings suggest that while LLM-based evaluators hold promise for scalable clinical assessment, their reliability remains dependent on both model design and the evaluation criteria being applied. This work provides insight into the strengths and limitations of judge models and offers a foundation for developing more robust automated evaluation methods in specialized domains.

2. Related Work

Prior research has shown that LLMs can generate unreliable responses in medical question-answering settings, where hallucinations and factual inaccuracies remain significant challenges (Reichenpfader et al., 2024; Yu et al., 2025). These limitations raise concerns about the safe deployment of LLMs in healthcare applications and highlight the need for robust evaluation methodologies.

To address this issue, a growing body of work has focused on systematically evaluating LLM-generated responses in medical domains. One approach relies on assessment by qualified biomedical experts. For instance, the CLEVER framework provides a structured methodology for trained medical professionals to evaluate LLM outputs (Kocaman et al., 2025). While expert evaluation is considered the gold standard, it is inherently time-

*These authors contributed equally.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 12–18
16 May 2026.

consuming and costly, limiting its scalability.

Alternatively, automatic evaluation metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) have been widely adopted to quantify the quality of generated text. However, these metrics primarily measure surface-level similarity and often fail to capture the nuanced clinical reasoning and domain-specific knowledge required in medical question answering. As a result, their effectiveness in evaluating medical responses remains limited.

Motivated by these shortcomings, recent work has explored the use of LLM-based evaluators under the "LLM-as-a-judge" paradigm. For example, Medical Eval Sphere introduces an open-source benchmark for long-form medical question answering based on real-world patient queries and expert-annotated reference responses (Hosseini et al., 2024). Similarly, MedThink-Bench presents a dataset of complex medical questions and incorporates an LLM-based judge to evaluate reasoning quality (Zhou et al., 2025). These efforts demonstrate the potential of automated, model-based evaluation frameworks in capturing more nuanced aspects of response quality.

Despite these advances, limited work has systematically examined how to train judge models that can reliably approximate expert evaluations, particularly with respect to the consistency of the underlying evaluation metrics. The impact of different training strategies and model combinations on metric reliability remains underexplored. In this work, we address this gap by leveraging the MEDIQA-EVAL dataset to study the performance of judge models across diverse learning paradigms. We further propose an ensemble-based approach and show that combining multiple models leads to more robust alignment with expert ratings than any single method.

3. Dataset and Task Description

This section outlines the dataset and the evaluation objectives of the MEDIQA-EVAL 2026 shared task. It details the composition and annotation schema of the dataset, with a formal definition of the evaluation metrics used to measure the effect between our system's predictions and expert judgments.

3.1. Dataset

The MEDIQA-EVAL 2026 shared task dataset is based on two multimodal medical visual question answering (VQA) collections: **DermaVQA** (dermatology cases from the MEDIQA-M3G challenge) and **WoundcareVQA** (wound care cases from the MEDIQA-WV 2025 challenge). These datasets consist of online medical posts, which include both textual descriptions and clinical images, paired with responses provided by medical professionals. Each clinical case is provided with gold-standard expert responses along with three distinct system-generated candidate responses.

Annotation Schema and Metrics To construct the ground-truth ratings, the system-generated responses were evaluated by human experts. There are 6 metric to consider.

- **Disagree_Flag:** if human experts disagree with an answer.
- **Factual-Accuracy:** if the answer contains all necessary information required to give to the patient based on the question present, rate from 0 and 1.
- **Relevance:** if everything in the response is medically accurate, rate from 0, 0.5, 1.0.
- **Completeness:** if the response covers all aspects of the user's query, rate from 0, 0.5, 1.0.
- **Writing-Style:** evaluates the linguistic quality, rate from 0, 0.5, 1.0.
- **Overall:** the overall score of the response, rate from 0, 0.5, 1.0.

Our work exclusively focuses on the **English (EN) sub-task**, utilizing the 2,898 data in the development set and 3,474 in the test set. The development set is the set we utilize to train our model, and the test set is the final evaluation set that computes the score on the leaderboard. The participants could not access the test set during the competition, though it was released after the competition for verification and reference.

3.2. Task

The MEDIQA-EVAL 2026 shared task evaluates the performance of the rating model using correlation with expert ratings. The objective is to produce predicted ratings that are the most correlated with expert-provided ratings. The final score is computed as the mean of Kendall, Spearman, and Pearson correlations according to the following formulas:

Define n to be the number of samples. $\hat{Y}_{1:n} \in R$ as the predicted scores and $Y_{1:n} \in R$ as the labeled scores for each metric.

Let

$$y_i, y_j \in Y, \quad \hat{y}_i, \hat{y}_j \in \hat{Y}, \quad i, j \in \{1, 2, \ldots, n\},\ i < j$$

Then, estimations of the correlations are:

$$\hat{\rho}_{pearson} = \frac{Cov(Y,\hat{Y})}{\hat{\sigma}_Y\, \hat{\sigma}_{\hat{Y}}}$$

$$\hat{\rho}_{spearman} = \frac{Cov(rank(Y), rank(\hat{Y}))}{\hat{\sigma}_{rank(Y)} \hat{\sigma}_{rank(\hat{Y})}}$$

$$\hat{\rho}_{kendall} = \frac{2}{n(n-1)} \sum_{i<j} sgn(y_i - y_j)\, sgn(\hat{y}_i - \hat{y}_j)$$

The final correlation is calculated as the mean of the three correlations:
$\hat{\rho} = \frac{\hat{\rho}_{pearson}+\hat{\rho}_{spearman}+\hat{\rho}_{kendall}}{3}$

4. Methodology

4.1. Few-Shot Learning

Our Few-Shot Learning method aims to help the judge LLM to understand the underlying ratings rules of the six metrics. We select Qwen3-30B-A3B as our base language model for its inference speed and text-only task performance(QwenTeam, 2025). There are 483 English samples in the development set of the task. To implement few-shot learning strategy, we perform 7 separate runs. In each run, we bootstrap 20 samples and feed them to the model. For each sample, the final output is computed by averaging the predictions from all 7 runs. The approach is shown in Fig 1.

4.1.1. Prompt Design

Prompt

You are an expert in dermatology and wound care. Think silently if needed .Your job is to give accurate ratings of the response generated by an LLM, which includes a diagnosis and a treatment. The response is an answer to a question from a patient asking about a disease that he has. You will be supplied with gold responses given by professional doctors. You should rate the LLM's response by comparing it very carefully with the gold responses and treat them as ground truth.

...<rating rules>...

I will give you a few examples with ratings. After that, you should provide your ratings to a sample, which contains only a query, a response and gold responses.

...<shots>...

The sample you need to rate:
<patient's query>
<LLM's response to the patient's query>
<gold responses to the patient's query>

For each sample, we construct our prompt with information containing the user's query, the LLM's response to the query, the gold responses, official rating rules presented by the MEDIQA-EVAL 2026 organizers, and the shots. We specifically instruct the rating LLM to compare carefully with the gold responses and deduce the underlying rating rules from the shots containing expert-labeled ratings. Using gold responses is a limitation of our approach, as it prevents direct application to datasets outside MEDIQA-EVAL. This design is intended to efficiently rate responses within this specific dataset.

4.1.2. Bootstrapped Few-Shot Learning

Although the rating LLM yields a decent result after few-shot learning with 20 shots, the rating LLM is still blind to other 483 training data . With only 20 shots, the rating LLM fails to see all possible combinations of metric scores. Therefore, we construct 7 conversations for each sample where each prompt contains bootstrapped shots from the training data, and the final output for the sample is the average of 7 LLM outputs from 7 conversations. We found that by bootstrapping shots multiple times and aggregating the final output by the mean consistently outperforms the model where the LLM is provided the shots for only once. We hypothesize that the ensemble of the 7 rating LLMs learns a better approximation of the distribution by sampling shots with replacement.

4.2. Fine-tuning BERT

4.2.1. Model Selection

PubMedBERT was selected as the base text encoder considering that it was pretrained on biomedical literature. Due to the same reason, *MedGemma27b-it* was employed for zero-shot image captioning. This effectively translates visual clinical information into textual descriptions, enabling the text-only BERT model to process multimodal information.

4.2.2. Targeted Upsampling

To decrease the severe class imbalance in the WoundcareVQA subset , a targeted upsampling mechanism was applied. Minority class instances (scores 0.0 and 0.5) were oversampled in the training folds to prevent the majority-class collapse.

4.2.3. Metric-Specific Input Construction

To accurately predict the six distinct evaluation metrics, it has naturally occurred that different metrics need different amount of information. Thus a metric-specific strategy was engineered.

For metrics like *writing-style*, *completeness* dependent primarily on linguistic quality, inputs were restricted strictly to the LLM response. The input sequence is constructed as the following: `[LLM Response] <response>`.

Conversely, for metrics like *relevance*, *factual-accuracy*, we constructed the inputs by concatenating the MedGemma27b-generated caption, query, LLM response. The input sequence is constructed as the following: `[IMG] <caption> [QUERY] <query> [LLM Response] <response>`.

4.2.4. Training and Inference Strategy

To fine-tune BERT for score generation, we adopt a tunable linear projection layer and apply a sigmoid function at the end to output a score from 0 to 1, as shown in Fig 1.

To maximize data utilization and prevent overfitting, a 5-fold cross-validation framework was employed. Models were optimized using cross-entropy loss over the three target classes. During inference, an internal probability averaging mechanism was utilized. This aggregation produces robust baseline predictions for the final system ensemble.

4.3. Retrieval Augmented Generation

Knowledge Base Construction: To ensure the factual accuracy of the evaluation, we constructed a domain-specific knowledge base utilizing 43 clinical guidelines focused on dermatology and wound care (World Health Organization, 2026; National Institute for Health and Care Excellence, 2026; NHS England, 2023; American Academy of Dermatology, 2026; Wound, Ostomy and Continence Nurses Society, 2026; Infectious Diseases Society of America, 2026; Wound Healing Society, 2026; American Physical Therapy Association, 2026). The majority of documents were systematically curated from authoritative international and professional repositories. They were retrieved from the official websites of leading medical institutions. Additionally, we incorporated evidence-based guidelines from specialized professional societies. We employed a sliding window strategy for text chunking, with a chunk size of 1024 tokens and an overlap of 100 tokens to preserve semantic continuity across boundaries.

Embedding and Indexing: We implemented a two-stage retrieval pipeline using the LlamaIndex framework (Liu, 2022) to retrieve relevant clinical evidence for a given gold response. The chunked text data was mapped into a high-dimensional vector space using the bge-m3 embedding model (Chen et al., 2024a). bge-m3 was selected for its robust performance in multilingual retrieval and support for dense retrieval tasks.

Retrieval and Reranking: For each patient query, we first performed an approximate nearest neighbor search to retrieve the top-10 candidate chunks. To further improve relevance and filter out noise, we applied a cross-encoder reranking step using bge-reranker-large (Xiao et al., 2023). This model re-scores the candidate chunks based on their fine-grained interaction with the query. The top-ranked chunks were then concatenated the top-3 chunks to form the context block denoted as <Relevant Clinical Guidelines> in the prompt.

4.4. Ensemble

As indicated by preliminary experiments on the development set, system performance varies across evaluation criteria, and no single modeling approach consistently achieves optimal performance across all metrics. The independent methods considered in this work—supervised BERT fine-tuning, Few-Shot LLM prompting, and Retrieval-Augmented Generation (RAG)—exhibit complementary strengths under different evaluation settings.

To leverage these complementary properties, we adopt an ensemble strategy for the final submission. Instead of employing a uniform or weighted aggregation scheme, we implement a dataset- and metric-specific selection mechanism. Concretely, for each combination of dataset and evaluation metric, the final prediction is derived exclusively from the model that achieves the highest correlation on the set. This design choice is motivated by empirical observations that model performance is highly dependent on both the task formulation and evaluation dimension.

- **Supervised BERT for Writing-Style:** The fine-tuned BERT model was exclusively selected for *writing-style* across both datasets. This confirms that supervised learning on domain-specific data remains the most effective method for capturing the stylistic consistency of medical responses.
- **RAG for Completeness:** The Retrieval-Augmented Generation (RAG) framework demonstrated superior performance on *relevance* for both datasets. This suggests that retrieving similar historical cases aids significantly in establishing the semantic alignment between the user query and the candidate response.
- **Few-Shot LLM for Factual Accuracy, Overall Ratings, Relevance and Disagree Flag:** The Large Language Model (LLM) proved most robust for high-level reasoning tasks, serving as the primary predictor for *factual-accuracy* and *overall* ratings. This underscores the necessity of leveraging the broad medical knowledge and zero-shot reasoning capabilities of LLMs for complex factual verification.

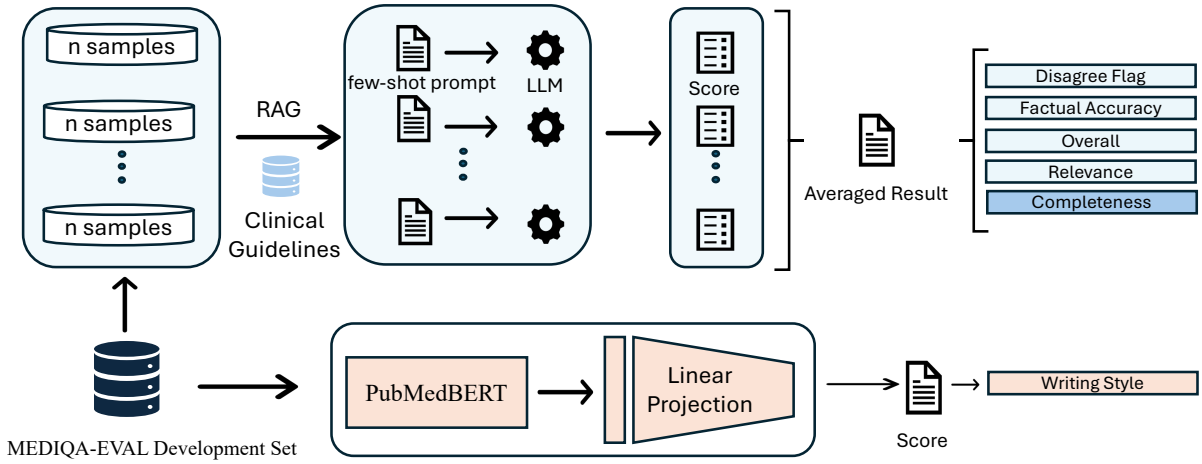

Figure 1: The proposed ensemble architecture for medical response evaluation.

Model	Disagree	Complete	Factual	Relevance	Style	Overall
Few-Shot	**0.48**	**0.39***	**0.44**	**0.57**	0.47	**0.47**
BERT	0.23	0.33	0.29	0.40	**0.52**	0.28
RAG	0.23	**0.38***	0.27	0.28	0.25	0.39
Ensemble	**0.48**	**0.39**	**0.44**	**0.57**	**0.52**	**0.47**

Table 1: Empirical result of each model on the development set.
*It can reasonably be assumed that RAG and few-shot learning perform equally well on this task. We chose RAG for metric 'completeness' on the test set.

By systematically exploiting these complementary strengths—fine-tuning for style, retrieval for relevance, and LLM reasoning for accuracy—our final system underscores the necessity of constructing an ensemble for complex medical evaluation tasks.

4.5. Results

The performance of our final ensemble system was evaluated on both the official MEDIQA-EVAL 2026 English development set and test set. The system, which integrates the strengths of supervised fine-tuning, retrieval-augmented generation, and few-shot learning, achieved a global aggregate mean correlation of **0.4816** (`ALL-en-ALL-mean`) on the test data.(Asma Ben Abacha, 2026)

Table 1 and Table 2 represent the detailed scores on development set and test set, respectively. The system demonstrated exceptional performance on the six metrics. This suggests that the ensemble strategy effectively synthesizes multimodal context and external knowledge to verify the semantic alignment and factual correctness of medical responses. It also indicates that the ensemble generalizes to data outside of the development set.

4.6. Error Analysis

To study the weakness of our model, we look at predictions that have a difference between the label larger than 0.5 on the development set. We found that poor predictions prevail at disagree flag and factual accuracy. A simple statistic showing where errors occur is shown in Table 3. These two metrics require solid knowledge of medical care and strong reasoning skills to detect the flaw in the response compared to other metrics. Further investigation is needed to understand the reason behind the large discrepancy of disagreeing a response between the judge model and human evaluators.

5. Conclusion

In this work, we show that an ensemble of few-shot learning, fine-tuned PubMedBERT, and RAG-enhanced prompting outperforms any individual model, with each component contributing distinct strengths across evaluation metrics. Specifically, fine-tuned PubMedBERT performs well on simpler classification tasks such as writing style, while LLM-based few-shot prompting is effective for metrics requiring medical reasoning, such as factual correctness. Incorporating retrieval-augmented information further improves performance on completeness. The consistent performance of our ensemble across both development and test sets suggests its robustness and generalizability.

Despite achieving strong overall results, the model shows weaker correlation on metrics such as disagree flag and writing style, and struggles with

Model	Disagree	Complete	Factual	Relevance	Style	Overall
Few-Shot	**0.36**	0.45	**0.54**	**0.60**	0.40	**0.55**
BERT	0.20	0.38	0.30	0.33	**0.40**	0.36
RAG	0.34	**0.45**	0.52	0.37	0.37	0.52
Ensemble	**0.36**	**0.45**	**0.54**	**0.60**	**0.40**	**0.55**

Table 2: Experiment on the contribution of each model on the test set, consistent with the findings on the development set

Metric	Count
Disagree_flag	70
Factual-Accuracy	12
Completeness	8
Overall	5
Relevance	2
Writing-Style	0

Table 3: Distribution of metrics for which error is higher than 0.5

aspects of factual accuracy. These limitations highlight the need for future work on improving medical reasoning capabilities and developing more reliable evaluation frameworks that better align with expert judgment.

Overall, our findings indicate that a carefully designed judge model can achieve meaningful alignment with expert ratings, though performance remains metric-dependent. Without appropriate design choices, individual methods may produce evaluations that are inconsistent with those of human experts.

6. Bibliographical References

American Academy of Dermatology. 2026. Aad guidelines of care. [EB/OL]. [2026-02-03].

American Physical Therapy Association. 2026. Apta clinical practice guidelines. [EB/OL]. [2026-02-03].

Wen-wai Yim Asma Ben Abacha. 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*. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

M. L. Chen et al. 2024b. Performance and risk of harm of a large language model on dermatology continuing medical education questions. *Journal of Investigative Dermatology*, 144(8):S25.

Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. 2025. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine*, 5(2):151–163.

Luca Corradini, Giulia Marcaccini, Isha Seth, Warren M. Rozen, Chiara Biagiotti, Roberto Cuomo, and Francesco R. Giardino. 2025. Ai vs. md: Benchmarking chatgpt and gemini for complex wound management. *Journal of Clinical Medicine*, 14(24):8825.

Pedram Hosseini, Jessica M Sin, Bing Ren, Bryceton G Thomas, Elnaz Nouri, Ali Farahanchi, and Saeed Hassanpour. 2024. A benchmark for long-form medical question answering. *arXiv preprint arXiv:2411.09834*.

Infectious Diseases Society of America. 2026. Idsa practice guidelines. [EB/OL]. [2026-02-03].

Veysel Kocaman, Murat Kaya, Adrian Feier, and David Talby. 2025. Clinical large language model evaluation by expert review (clever): Framework development and validation. *JMIR AI*, 4:e72153.

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.

Jerry Liu. 2022. LlamaIndex.

National Institute for Health and Care Excellence. 2026. Nice guidelines: Evidence-based recommendations for health and care. [EB/OL]. [2026-02-03].

NHS England. 2023. Wound care information standard. [EB/OL]. [2026-02-03].

QwenTeam. 2025. Qwen3 technical report.

Daniel Reichenpfader, Peter Rösslhuemer, and Kerstin Denecke. 2024. Large language model-based evaluation of medical question answering systems: Algorithm development and case study. *Studies in Health Technology and Informatics*, 313:22–27.

World Health Organization. 2026. World health organization: Guidelines and technical documents. [EB/OL]. [2026-02-03].

Wound Healing Society. 2026. Wound healing society clinical resources and guidelines. [EB/OL]. [2026-02-03].

Wound, Ostomy and Continence Nurses Society. 2026. Wocn society: Clinical guidelines and resources. [EB/OL]. [2026-02-03].

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

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. 2025a. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025b. MORQA: benchmarking evaluation metrics for medical open-ended question answering. volume abs/2509.12405.

Ww. Yim, Y. Fu, Z. Sun, A. B. Abacha, M. Yetisgen, and F. Xia. 2024. Dermavqa: A multilingual visual question answering dataset for dermatology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15005 of *Lecture Notes in Computer Science*, Cham. Springer.

E Yu, X Chu, W Zhang, X Meng, Y Yang, X Ji, and C Wu. 2025. Large language models in medicine: Applications, challenges, and future directions. *International Journal of Medical Sciences*, 22(11):2792–2801.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Sheng Zhou, Wei Xie, Jian Li, Zhi Zhan, Ming Song, Hui Yang, Carlos Espinoza, Laura Welton, Xiang Mai, Yan Jin, Zhen Xu, Young-Hoon Chung, Yi Xing, Ming-Hsiu Tsai, Eric Schaffer, Yi Shi, Nan Liu, Zhi Liu, and Rong Zhang. 2025. Automating expert-level medical reasoning evaluation of large language models. *NPJ Digital Medicine*, 9(1):34.