# SloCal-Net at MEDIQA-Eval 2026: Investigating the Impact of Reasoning and External Context on Medical Answer Grading

**Primoz Kocbek**[1,2], **Valentina Carbonari**[3], **Pierangelo Veltri**[4], **Pietro Hiram Guzzi**[3], **Gregor Stiglic**[1,5]

[1] University of Maribor, Faculty of Health Sciences, Žitna ulica 15, 2000 Maribor, Slovenia, {primoz.kocbek, gregor.stiglic}@um.si,

[2] University of Ljubljana, Faculty of Medicine, Vrazov trg 2, 1000 Ljubljana, Slovenia,

[3] Magna Graecia University of Catanzaro, Department of Surgical and Medical Sciences, Viale Europa, 88100 Catanzaro, Italy, {valentina.carbonari, hguzzi}@unicz.it

[4] University of Calabria, DIMES, Via P. Bucci, 87036 Rende, Italy, pierangelo.veltri@unical.it,

[5] Usher Institute, The University of Edinburgh, Edinburgh EH16 4UX, UK

**Abstract**

Automated evaluation of multimodal medical answers is essential for scalable safety assessment, yet it remains difficult to align automatic scores with expert judgment across languages and image modalities. We describe SloCal-Net's systems for the MEDIQA-EVAL 2026 shared task, framing evaluation as rubric-conditioned multimodal judging: the judge receives the question, image(s), candidate answer, and task-specific criteria, and outputs criterion-level scores and an overall rating. Evidence retrieval was initialized using ChatGPT Deep Research, producing a 25-document clinical corpus used for lightweight retrieval-augmented grounding. On the official leaderboard, our best submission (GPT-5-mini with web search and structured scoring) achieved Pearson correlations of 0.466 on English and 0.260 on Chinese expert ratings. In post-competition experiments with open-source judges, the best English Pearson reached 0.272 with GLM-4.6V and 0.212 with Qwen3-VL-30B-Thinking, while Chinese correlations were lower, highlighting remaining gaps in multilingual calibration and image–text grounding.

**Keywords:** multimodal evaluation, clinical NLP, LLM-as-a-judge

## 1. Introduction

The integration of large multimodal language models (LLMs) into clinical workflows has led to a paradigm shift in the generation and assessment of medical responses. However, ensuring reliability, safety, and clinical validity remains a major challenge. Conventional evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) primarily measure lexical overlap and often fail to reflect medical correctness or appropriateness. Similarly, general-purpose vision–language models such as CLIP (Radford et al., 2021) lack the domain specificity required for nuanced interpretation of medical images. In practice, evaluation systems must perform cross-modal reasoning—aligning written responses with visual clinical evidence such as dermatological images or diagnostic scans—while accounting for the fact that multiple expert answers may be clinically acceptable (Yim et al., 2025).

Historically, automated medical evaluation relied on n-gram–based metrics, which correlate poorly with clinician judgment in open-ended settings (Asma Ben Abacha, 2026). This limitation has motivated a shift toward semantic and reasoning-based evaluation frameworks. More recently, the *LLM-as-a-judge* paradigm has emerged, where strong models are prompted to score candidate outputs against structured clinical rubrics (Zheng et al., 2023). While effective for text-only tasks, purely textual judges struggle in multimodal contexts, where visual evidence must be verified against generated claims (Anonymous, 2025). Shared tasks such as MEDIQA-EVAL therefore require evaluators to detect inconsistencies between patient images and candidate responses (Yim et al., 2025b).

Parallel advances in domain-specific multimodal modeling further shape this landscape. Generalist vision–language encoders often fail to capture fine-grained dermatological distinctions (Yan et al., 2025), motivating specialized models such as DermLIP, trained on the Derm1M dataset of clinically aligned dermatology image–text pairs (Yan et al., 2025). Similarly, MedGemma integrates medical vision encoders optimized for radiology and pathology (Sellergren et al., 2025). These models illustrate the growing importance of domain-specific representations for clinically grounded evaluation.

At the same time, newer reasoning-oriented architectures emphasize inference-time deliberation and structured decision-making. Systems such as the GPT-5 family of models incorporate internal planning mechanisms to improve complex multi-



*Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026*, pages 235–243

modal reasoning (Wang et al., 2025). Retrieval-augmented and agentic workflows further attempt to reduce hallucinations by grounding predictions in external evidence (OpenAI, 2025). Platform-level efforts, including large-scale clinical AI ecosystems and evaluation frameworks such as ChatGPT Health highlight the increasing emphasis on safety, alignment, and real-world deployment constraints.

Within this context, the SloCal-Net team approach adopts a pragmatic LLM-judge strategy for multimodal clinical evaluation. We develop prompt-based judges and study two design dimensions that directly affect performance: (i) per-dimension scoring versus structured per-case evaluation, and (ii) the role of retrieval-augmented grounding. We report official shared-task results and conduct additional analyses with open-weight multimodal judges to assess the remaining performance gap between proprietary and open systems (Asma Ben Abacha, 2026).

## 2. Task and Data

The MEDIQA-EVAL 2026 shared task addresses this need by providing a benchmark for the automatic evaluation of multimodal medical responses, building on recent work on evaluating medical open-ended question answering such as MORQA (Yim et al., 2025a). The competition utilizes two primary multimodal datasets cantaining both image and textual data, i.e. WoundCare (Yim et al., 2025b), which focuses on wound management, and IIYI (Yim et al., 2024), which covers broader clinical cases. A series of questions were posed to three distinct models on the various cases proposed by the datasets, and the responses were then compared. The primary objective is to automatically evaluate the responses provided by the three systems, which are further subdivided into two subtasks: one in English (Subtask 1) and one in Chinese (Subtask 2). In the English subtask, systems are tasked with predicting six distinct criteria: disagreement flag, completeness, factual accuracy, relevance, writing style, and overall. With regard to the Chinese subtask, the evaluation is centered on two primary components: factual consistency and writing style. The scores ranged from 0 to 1 for each metric, where 0 indicates a terrible response (incorrect or/and harmful; may lead to grievous misinformation or dangerous outcomes; should never be shown to the patient), 0.5 indicates a response that is relevant and will not necessary cause harm but misses critical items or contains incorrect information (should not be shown to the patient before correction), and 1 indicates a complete and accurate response that can be safely presented to a patient. Note that disagreement flag was a binary choice.

The predictions were evaluated against the human scores using Pearson, Spearman, and Kendall's Tau correlations (Shaqiri et al., 2023).

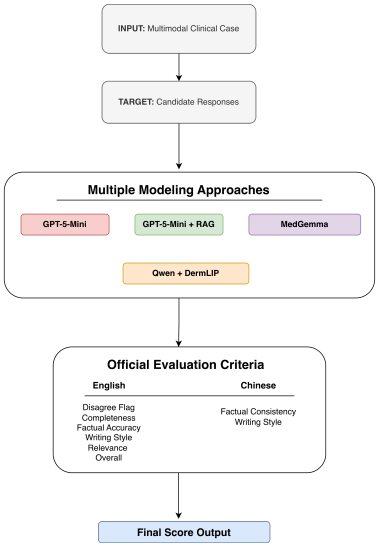


Figure 1: SloCal-Net workflow for multimodal evaluation of medical answers.

We evaluate several judging pipelines (Figure 1): generic reasoning with GPT-5-mini (Papineni et al., 2002) (zero-shot and RAG), specialized medical reasoning with MedGemma (Sellergren et al., 2025) and other state-of-the-art open source models (Section 5.5), and a local multimodal pipeline that combines DermLIP for visual features, KeyBERT (Grootendorst, 2020) and SBERT (Reimers and Gurevych, 2019) for semantic text processing, and Qwen-2.5 as the reasoning backbone.

## 3. Data and Resources

This study uses two publicly available multimodal clinical Visual Question Answering (VQA) benchmarks: **DermaVQA** (Yim et al., 2024) and **WoundcareVQA** (Yim et al., 2025b). Both datasets pair clinical images with natural-language questions and expert-written answers, but they differ in domain (dermatology vs. wound care), scale, and available metadata (Table 1). DermaVQA combines Chinese cases from IIYI with English cases from Reddit, and includes a subset with patient metadata (age/sex/Fitzpatrick skin type), while WoundcareVQA focuses on wound assessment with rich structured labels (e.g., anatomic location, wound type and stage, drainage, and infection status) and expert responses authored by US clinicians.

DermaVQA and WoundcareVQA are multilingual, image-grounded QA resources designed to reflect real patient–provider interactions (Yim et al., 2024, 2025b). DermaVQA centers on dermatology, com-

| Feature | DermaVQA | WoundcareVQA |
|---|---|---|
| Medical Domain | Dermatology | Mainly Wound Care |
| Total Images | 3,434 | 748 |
| Total Cases/QA | 1,488 question-answer pairs | 477 cases (768 expert responses) |
| Data Sources | IIYI (Chinese) and Reddit (English) | Tieba and Zhidao (Chinese translated to English) |
| Clinical Metadata | Age, Sex, and Fitzpatrick Skin Type (Reddit subset only) | 41 anatomic locations, 8 wound types, 6 thickness stages, tissue color, drainage, and infection |
| Expert Involvement | Medical translators (IIYI) and professional dermatologists (Reddit) | Practicing US medical doctors (Surgeon, ER Physician, ER Resident) |

Table 1: Comparison between DermaVQA and WoundcareVQA datasets.

bining Chinese forum cases (IIYI) with English Reddit questions; WoundcareVQA targets wound assessment and triage, with expert responses and extensive structured metadata to support fine-grained clinical evaluation. A concise comparison of their domains, sizes, data sources, and metadata is provided in Table 1.

## 4. Tools Overview

The experimental setup combines multimodal reasoning models, domain-specific vision encoders, and semantic text-processing components. DermLIP (Yan et al., 2025) provides dermatology-oriented visual representations, GPT-5-mini (Wang et al., 2025) and MedGemma act as multimodal clinical judges, SBERT (Reimers and Gurevych, 2019) and KeyBERT (Grootendorst, 2020)support semantic alignment and keyword extraction, and Qwen models serve as multilingual reasoning backbones. Each component addresses a distinct stage of the evaluation pipeline, enabling controlled comparison across architectures.

- **GPT-5-mini** : A compact multimodal reasoning model used as the primary evaluation engine. It demonstrates strong performance on medical knowledge and multimodal QA benchmarks, supporting integrated interpretation of textual and visual clinical evidence.
- **Derm1M and DermLIP** : Derm1M is a large dermatology vision–language dataset aligned with clinical ontologies. DermLIP models trained on this resource enable fine-grained visual feature extraction and improved cross-modal retrieval in dermatological settings.
- **Sentence-BERT (SBERT)**: Produces sentence-level embeddings to estimate the semantic similarity between each candidate and a consensus clinical context—dynamically extracted from the pool of candidates via KeyBERT—to evaluate factual alignment.
- **KeyBERT** : Extracts clinically relevant keywords using contextual embeddings, aiding relevance and completeness assessment.
- **Qwen models** (Qwen Team, 2024; Team, 2024): Multimodal and multilingual transformer models used as reasoning backbones in the local pipeline, enabling structured inference across text and image inputs.

## 5. Methodology

Given the complexity of the WoundCare and IIYI datasets, we explored multiple judging strategies. We first tested RAG to provide external evidence to the judge at scoring time. We then enforced structured outputs so each model returned per-criterion scores in a consistent format. Next, we evaluated a fully local multimodal pipeline based on Qwen and DermLIP. Finally, we compared against MedGemma as a domain-specialized multimodal judge as well as general open-source models, such as Qwen3-VL and GLM-4.6V. Across all settings, the goal is to predict scores that reflect the judgment of clinical experts on a discrete scale $\{0, 0.5, 1.0\}$.

To evaluate the impact of these architectural choices, we configured our judging systems into five distinct experimental runs reported in Table 3:

- **GPT-5-mini WS**: A baseline multimodal judge using a web-search tool for real-time information retrieval without a static indexed corpus.
- **GPT-5-mini WS RAG**: The baseline system augmented with our 25-document clinical evidence corpus for grounded retrieval-augmented generation.
- **GPT-5-mini WS Struc**: The baseline system utilizing Pydantic-enforced schemas to generate rigid, multi-dimensional JSON scores.
- **GPT-5-mini WS RAG Struc**: A composite configuration combining both the static RAG corpus and structured output constraints.

- **DermLIP Qwen2**: Our fully local pipeline integrating specialized dermatological embeddings with a Qwen reasoning backbone.

Within these configurations, 'Web Search' (WS) refers to the model's native agentic ability to query live clinical resources, whereas 'RAG' specifically refers to grounding within our curated 279-page clinical document index.

## 5.1. Retrieval-Augmented Judging (RAG)

In the RAG setting, we augment the judge with external evidence so that scoring is explicitly grounded in verifiable background knowledge rather than relying solely on the model's parametric memory. We constructed an evidence corpus using an agentic web-search workflow (ChatGPT Deep Research) that (i) analyzed the official task website and evaluation protocol, and (ii) searched the literature for high-quality resources on multimodal medical reasoning, medical image interpretation (dermatology and wound care), and reliable evaluation methodologies for open-ended QA. The Deep Research prompt we used was:

> I need a comprehensive background knowledge for this competition https://sites.google.com/view/mediqa2026/mediqa-eval Analyse all the data from the website and search the web for resources that can the best ground a model, note that this mean both understanding medical images or how to interpret them and provided text. At the end provide 25 hyperlinks to the best knowledge to get a grounded understanding and how to best evaluate a statement. Be very scientific and precise. The link need have access to a pdf file, so double check it, since it will be use a additional context in a RAG.

## 5.2. Structured Output and Scoring

To assess an entire clinical encounter in a single multimodal inference step, we used prompts with Pydantic-enforced schemas. The model is required to generate: (i) **Visual Synthesis**, a detailed characterization of pathological findings in the provided images; (ii) **Integrated Reasoning**, a clinical chain-of-thought (CoT) that synthesizes visual evidence, case history, and expert references; and (iii) **Multi-Dimensional Scoring**, a structured JSON object containing discrete scores (0.0, 0.5, 1.0) for all required metrics (and, when applicable, a Disagree Flag). This structured format enables consistent aggregation and evaluation across judge models.

## 5.3. Qwen and DermLIP Pipeline

The proposed architecture, developed to test the dataset proposed for the competition, integrates visual and textual information through different modules, with the reasoning process guided by a Large

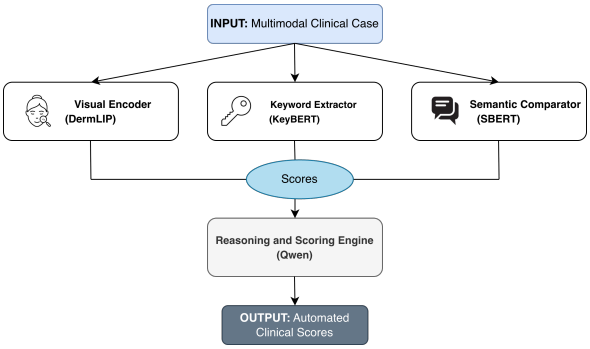


Figure 2: SloCal-Net workflow for Qwen and DermLIP.

Language Model. As shown in Figure reffig:workflow1, the evaluation process consists of three main stages: feature extraction, context integration, and final scoring. Given the specificity of the dermatological domain, clinical images are processed using DermLIP. This model, pre-trained on medical datasets, allows the extraction of visual embeddings that capture critical pathological details, overcoming the limitations of generalist visual models. The textual component is analysed along two lines:

- *Clinical Relevance:* We use KeyBERT to extract essential medical keywords from the entire pool of candidate responses, ensuring that key concepts are not overlooked.
- *Semantic Coherence:* We employ SBERT to calculate the vector similarity between the candidate's generated response and the reference, providing a quantitative measure of semantic alignment.

The system is configured to produce discrete scores in the range $\{0, 0.5, 1.0\}$ for different qualitative dimensions; only for Factual Accuracy are the values not discretised. For the English language subtask, the following are evaluated: Completeness, Factual Accuracy, Relevance, Writing Style, Overall Quality, and a Disagree Flag for clinical safety. For the Chinese language subtask, the system focuses on *Factual Consistency* and *Writing Style*.

For the Factual Accuracy metric, Qwen integrates the visual features provided by DermLIP into the decision-making process, allowing for clinical accuracy verification anchored to photographic evidence of the case. For the remaining metrics (such as Completeness, Relevance and Writing Style), Qwen acts as an informed zero-shot classifier, analysing the consistency between the keywords extracted by KeyBERT, the semantic similarity calculated by SBERT and the content of the candidate response to assign the final scores. The

Qwen model therefore constitutes the central reasoning unit of the pipeline and is used via a prompting technique; specifically, prompt-based assessment is reserved for qualitative aspects, where the model acts as a 'senior clinician' to assign discrete scores. In this context, Qwen evaluates the candidate's response by checking its clinical consistency and the use of professional terminology.

### 5.4. GPT-5-mini

We used GPT-5-mini as our primary multimodal judge, both in a zero-shot setup and with retrieval-augmented generation (RAG). In the zero-shot configuration, the model directly scored each response according to the official rubric criteria. In the RAG configuration, the judge was additionally grounded with a small set of retrieved external documents (produced via a web-search tool) intended to provide task-relevant medical context and reduce unsupported claims. This setup follows common patterns in recent evaluation competitions that leverage rubric-driven *LLM-as-a-judge* scoring and study grounding to improve factuality and calibration (Zheng et al., 2023; Anonymous, 2025; Asma Ben Abacha, 2026). To ensure reproducibility, the same vision-language model utilized as the primary scoring engine (GPT-5-mini) was also employed to generate the image-aware summaries used for constructing the retrieval queries.

### 5.5. Open-source models

We carried out a **post-competition** evaluation with the open-source multimodal models reported in Table 4:

- **GLM-4.6V-109B-Instruct** (Zhipu AI, 2025). A large-scale vision–language model from the GLM lineage designed for multimodal reasoning across images and documents. It supports instruction-following and visual grounding for tasks such as medical VQA and report interpretation.
- **Qwen3-VL-30B-Thinking** (Qwen Team, 2025). A reasoning-focused multimodal model optimized for structured inference over visual and textual inputs, building on the Qwen-VL framework for image understanding, grounding, and multimodal dialogue (Qwen Team, 2023). These capabilities enable diagnostic-style reasoning workflows combining imaging and clinical context.
- **Qwen3-VL-32B-Instruct** (Qwen Team, 2025). Instruction-tuned multimodal model for document parsing, visual grounding, and image-conditioned response generation, aligned with the general multimodal capabilities established in the Qwen-VL technical report (Qwen Team, 2023), including applications to medical image–text interpretation tasks.
- **MedGemma-27B-it** (Sellergren et al., 2025). A medical-domain vision–language model designed for clinical reasoning across radiology and pathology imagery. It supports clinical question answering and report-style generation.
- **MedGemma-1.5–4B-it** (Google, 2026). A compact multimodal clinical model optimized for deployment and fine-tuning; smaller medical vision–language backbones have been used for medical reasoning in resource-constrained settings.

We ran these models in the same experimental setup as our other judges: each model received the image, question, candidate answer, and rubric definition (and, in the RAG variant, the retrieved evidence context) and then produced per-criterion scores. These experiments were conducted after the official submission deadline and are therefore reported as supplementary analyses rather than leaderboard submissions. All open-source models were served locally on an HGX $8\times$H200 server via vllm; we used a single H200 GPU for all models except GLM-4.6V-109B-Instruct, which was run on two H200 GPUs.

## 6. Results and Discussion

### 6.1. Official shared-task results

| Team | Score(en) | Score(zh) |
|---|---|---|
| MEDIQA | 0.5551 | 0.2767 |
| SUAT-BMI | 0.5450 | / |
| MedAware | 0.5261 | 0.2505 |
| **SloCal-Net (Ours)** | 0.4656 | 0.2596 |
| BDI | 0.4392 | 0.2524 |
| hgkai26 | 0.2444 | / |

Table 2: Official leaderboard results.

The official leaderboard performance (Table 2), measured as Pearson correlation between expert ratings and automatic scores for English and Chinese. Our primary submission (GPT-5-mini with web search and rubric-conditioned scoring) achieved correlations of 0.4656 (English) and 0.2596 (Chinese). These results suggest that general-purpose multimodal reasoning models can approximate expert judgment in rubric-driven grading tasks.

However, a performance gap remains relative to the top-ranked systems, particularly for English.

| Strategy | RAG | Overall mean (en) | Overall mean (zh) | Latency (s) | Total Tokens |
|---|---|---|---|---|---|
| GPT-5-mini WS | No | **0.4898** | 0.2555 | 30.64 | 3,155 |
| GPT-5-mini WS RAG | Yes | 0.4656 | **0.2596** | 26.93 | 3,805 |
| GPT-5-mini WS Struc | No | 0.3021 | 0.1649 | 28.47 | 3,092 |
| GPT-5-mini WS RAG Struc | Yes | 0.2436 | 0.1628 | 27.23 | 3,864 |
| DermLIP Qwen2 | No | 0.2428 | 0.1678 | 16.10 | 614 |

Legend: WS = web-search tool; Struc = structured output.

Table 3: Submitted runs with latency and total tokens used.

This suggests that while strong reasoning capacity is necessary, calibration to task-specific rubrics and multimodal grounding is equally critical. The lower Chinese performance reflects broader challenges in multilingual alignment, including differences in rubric interpretation, linguistic ambiguity, and variability in clinical terminology. This performance gap is primarily driven by the English-centric bias in foundational medical pre-training data, which results in less reliable cross-modal grounding for Chinese clinical narratives. Furthermore, the inherent difficulty of translating highly nuanced clinical evaluation rubrics introduces subjective variance that disrupts the models' alignment with expert judgments.

### 6.2. Ablations: retrieval and output-structure effects

Looking at controlled ablations (Table 3), where we isolate the impact of retrieval and structured outputs on GPT-5-mini. We firstly observe, that retrieval provided a modest improvement for Chinese (0.2555 $\rightarrow$ 0.2596) but reduced performance for English (0.4898 $\rightarrow$ 0.4656). The deterioration in performance, in this case, may be linked to semantic noise and 'semantic bleeding', i.e. the inclusion of external documents may introduce irrelevant information that distracts the model.

A post-hoc analysis revealed that approximately 15% of retrieved passages contained 'semantic noise', medically accurate but contextually irrelevant protocols, which triggered 'semantic bleeding' and a 2.4% decrease in English Pearson correlation. Conversely, this same external context likely provided a necessary terminological bridge in the Chinese subtask, where retrieval noise was outweighed by the benefits of improved multilingual grounding. In multimodal contexts (text + image), this translates into "semantic bleeding"(Brown et al., 2025). The poor performance of specialised workflows (0.24) reflects a structural challenge in the field of dermatology linked, for example, to the so-called "granularity gap": models may be excellent at binary screening but fail dramatically when it comes to identifying specific subclasses (Yuceyalcin et al., 2026).

Furthermore, in the semantic space of the models, visually similar lesions tend to overlap. This overlap makes it almost impossible for workflows such as DermLip + Qwen to correctly separate pathologies that, although biologically different, appear similar to AI. The low score of the DermLip-based workflow is further justified by the technical limitations of CLIP-based models, such as text and token number limitations (Yan et al., 2025).

Our second observation is that structured output constraints reduced correlation for both languages. While structured scoring improves interpretability and reproducibility, it appears to over-constrain the judge and limit flexible reasoning across modalities. This might indicate trade-off between evaluation transparency and predictive alignment.

Latency and token consumption further reveal a practical trade-off: retrieval increased computational cost without consistent performance gains, suggesting that evidence selection and filtering are critical design factors.

### 6.3. Modular Ablation: Impact of DermLIP and KeyBERT-SBERT Context

To assess the contribution of each module used in the hybrid evaluation pipeline (Qwen + DermLIP + KeyBERT), we conducted a controlled ablation study on the validation set; the results for Factual Accuracy are shown in Table 5. The results show that the complete pipeline provides the most accurate alignment with expert evaluations. The most significant drop in performance occurred when the *Clinical Context* module was removed (**No-KeyBERT**). Similarly, the exclusion of specialised visual features (**No-DermLIP**) confirms that, whilst generic models such as Qwen possess strong reasoning capabilities, they lack the diagnostic sensitivity of expert-level dermatological latent spaces.

### 6.4. Post-competition evaluation of open-source judges

Post-competition experiments evaluated open-weight multimodal models under identical prompting conditions (Table 4). The strongest English

| Model | RAG | Overall mean (en) | Overall mean (zh) | Latency (s) | Total Tokens |
|---|---|---|---|---|---|
| GLM-4.6V-109B-Instruct (Zhipu AI, 2025) | Yes | 0.2716 | 0.2168 | 272.83 | 8,044 |
| | No | 0.2512 | 0.2617 | 179.51 | 3,155 |
| MedGemma-27B-it (Sellergren et al., 2025) | Yes | 0.2559 | 0.1789 | 44.57 | 5,568 |
| | No | 0.1729 | 0.1613 | 34.54 | 2,319 |
| Qwen3-VL-30B-Thinking (Qwen Team, 2025) | Yes | 0.2117 | 0.1533 | 45.55 | 6,289 |
| | No | 0.2045 | 0.1466 | 21.72 | 2,509 |
| Qwen3-VL-32B-Instruct (Qwen Team, 2025) | Yes | 0.1930 | 0.1616 | 47.27 | 6,137 |
| | No | 0.1875 | 0.1484 | 33.64 | 2,751 |
| MedGemma-1.5-4b-it (Sellergren et al., 2025) | No | 0.0899 | 0.0417 | 43.27 | 2,302 |
| | Yes | 0.0368 | 0.0821 | 46.03 | 5,543 |

Table 4: Post-competition analysis of open-source models.

| Configuration | Pear. ($r$) | Spear. ($\rho$) | Kend. ($\tau$) |
|---|---|---|---|
| **Full (Qwen+DermLIP+KB)** | **0.2188** | **0.2294** | **0.1810** |
| No-DermLIP (Vision Qwen) | 0.1916 | 0.1934 | 0.1520 |
| No-KeyBERT (No Context) | 0.1365 | 0.1519 | 0.1200 |

Table 5: Ablation results for the Qwen+DermLIP scoring pipeline.

performance was observed for GLM-4.6V (0.27 Pearson), followed by MedGemma and Qwen3-VL variants. Performance differences across models correlate strongly with scale and multimodal pretraining depth.

The performance lags significantly proprietary models, such as GTP-5-mini, indicating limited multilingual calibration and weaker rubric grounding. Smaller models showed even larger degradation, suggesting insufficient capacity for joint reasoning across image, language, and evaluation criteria.

These findings indicate that open-source multimodal judges remain competitive for exploratory or resource-constrained settings but still lag behind proprietary models in reliability and cross-lingual robustness.

Future work will focus on adaptive evidence selection strategies, and improved multimodal grounding for multi-image and low-quality clinical inputs.

### 6.5. Limitations

The proposed evaluation framework presents three primary limitations. First, system efficacy relies heavily on the quality of the RAG knowledge base, where poor retrieval directly degrades the judge's diagnostic accuracy. Second, the multimodal judge exhibits high prompt sensitivity, meaning subtle lexical variations can significantly alter its interpretation of complex clinical rubrics. Finally, deploying state-of-the-art proprietary models like GPT-5-mini incurs substantial token costs, creating a practical financial impediment to scaling the framework for high-volume clinical analysis.

## 7. Conclusions

We presented SloCal-Net's participation in the MEDIQA-EVAL 2026 shared task, treating automated evaluation as rubric-conditioned multimodal judging with optional retrieval grounding. Our best system (GPT-5-mini with web search and structured scoring) achieved Pearson correlations of 0.466 (English) and 0.260 (Chinese) against expert ratings. Post-competition tests show that open-weight multimodal judges can approach the English leaderboard performance (up to 0.402 Pearson) but exhibit a larger drop on Chinese, suggesting that multilingual calibration and rubric interpretation remain key bottlenecks. Overall, retrieval and structured outputs can help, but they require careful prompt design to avoid over-constraining the judge. Future work will focus on multilingual rubric normalization, evidence selection, and robustness to multi-image cases and varying image quality.

## 8. Acknowledgements

This work was supported by European Union under Horizon Europe [grant number 101159018] and EuroHPC JU [grant number 101101903]; Slovenian Research Agency [grant number N3-0307, GC-0001]. Project Unveiling RNAbased Therapeutic Targets for PostTransplant Nephropathy (SINNephro RNA) PNRR Next Generation EU CN00000041 National Center for Gene Therapy and Drugs based on RNA Technology. National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.5, project 'RAISE – Robotics and AI for Socio-economic Empowerment' (ECS00000035) under the project 'Gestione e Ottimizzazione di Risorse Ospedaliere attraverso Analisi Dati, Logic Programming e Digital Twin (GOLD)', CUP H53C24000400006.

## 9. Bibliographical References


Anonymous. 2025. Healmqa: A healthcare multimodal question answering dataset for benchmarking large language models. OpenReview preprint (under review at ICLR 2026). Accessed: 2026-02-17.

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.

Andrew Brown, Barry Devereux, and Muhammad Roman. 2025. Retrieval-augmented generation (rag): A systematic literature review of techniques, metrics, challenges, and future directions. *Big Data and Cognitive Computing*.

Google. 2026. MedGemma 1.5-4b-it (model card). Hugging Face model card. Accessed: 2026-02-17.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

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

OpenAI. 2025. Deep research system card. System card. Accessed: 2026-02-17.

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.

Qwen Team. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint. ArXiv:2308.12966.

Qwen Team. 2024. Qwen2 technical report. arXiv preprint. ArXiv:2407.10671.

Qwen Team. 2025. Qwen3-vl technical report. arXiv preprint. ArXiv:2511.21631.

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. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763.

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 (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report.

Mirlinda Shaqiri, Teuta ILJAZI, Lazim KAMBERI, and Rushadije RAMANI-HALILI. 2023. Differences between the correlation coefficients pearson, kendall and spearman. *Journal of Natural Sciences and Mathematics of UT*, 8(15-16):392–397.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. 2025. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224*.

Siyuan Yan, Ming Hu, Yiwen Jiang, Xieji Li, Hao Fei, Philipp Tschandl, Harald Kittler, and Zongyuan Ge. 2025. Derm1m: A million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12681–12690.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2025. Morqa: Benchmarking evaluation metrics for medical open-ended question answering. arXiv preprint. ArXiv:2509.12405.

Furkan Yuceyalcin, Abdurrahim Yilmaz, and Burak Temelkuran. 2026. A hierarchical benchmark of foundation models for dermatology. *arXiv preprint arXiv:2601.12382*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Zhipu AI. 2025. Glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable vision-language models. arXiv preprint. ArXiv:2507.01006.

## 10. Language Resource References


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. volume abs/2509.12405.

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*, 161:104639. Language resource / dataset paper.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. 2024. DermaVQA: A multilingual visual question answering dataset for dermatology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 209–219. Springer. Language resource / dataset paper.


# A. Appendices

## A.1. LLM Judge Prompts

This appendix details the exact system and user prompts used to instruct the LLMs for rubric-conditioned multimodal judging in both the English and Chinese subtasks.

### A.1.1. English Evaluation Prompts

**System Prompt:**

```
You are a Senior Clinical Evaluator. The identified body part in the image is: {body_part}. Perform a rigorous EXTREMELY CONCISE assessment.
AVOID FLUFF and filler text.

Evaluation Protocol (MANDATORY):
1. **Visual Synthesis**: Concise, high-yield description of pathology and landmarks, focusing on the {body_part} region.
2. **Clinical Reasoning**: Brief, step-by-step analysis comparing Candidate vs. Gold Standard. Focus strictly Safety (red flags), Accuracy, and Actionability.
3. **Structured Scoring**: Assign scores (0.0, 0.5, 1.0). All justifications MUST be under 50 words.

Scoring Rubric:
- **completeness**: 1.0 (Full/Correct), 0.5 (Partial), 0.0 (Incorrect/Missing).
- **factual-accuracy**: 1.0 (Accurate), 0.5 (Minor errors), 0.0 (Wrong).
- **relevance**: 1.0 (Relevant), 0.5 (Partial), 0.0 (Irrelevant).
- **writing-style**: Professionalism, clarity, and conciseness.
- **overall**: Summary of utility and safety.
- **disagree_flag**: (0.0 or 1.0) 1.0 if advice contradicts expert.
```

**User Prompt:**

```
### CLINICAL CASE
Title: {title}
Content: {content}

### RAG CONTEXT
{rag_ctx}

### EXPERT REFERENCE (GOLD STANDARD)
{ref_text}

### CANDIDATE ANSWER TO EVALUATE
{candidate}

Provide your analysis and scores in structured JSON. You MUST provide a score for ALL metrics defined in the schema.
```

### A.1.2. Chinese Evaluation Prompts

**System Prompt:**

```
您是资深医学专家评估员。图像中识别出的身体部位：您是资深医学专家评估员。图像中识别出的身体部位是：{body_part}。请进行严谨且极其简洁的评估，严禁废话和冗长描述。

评估流程：
视觉分析：结合已识别的部位（），简洁、高价值地1. **视觉分析**：结合已识别的部位（{body_part}），简洁、高价值地描述图像病理特征与解剖标志。
临床推理：简明扼要地对比专家标准，分析安全性（危险信号）与确。2. **临床推理**：简明扼要地对比专家标准，分析安全性（危险信号）与准确性。
结构化评分：为每个指标提供或的，理由必须在字以内。3. **结构化评分**：为每个指标提供 0.0, 0.5, 或 1.0 的评分，理由必须在 50 字以内。

评分细则：
- **factual-consistency-wgold**：
  - 1.0：完整且与至少一个参考回复事实相符。
  - 0.5：不完整但与至少一个参考回复事实相符。
  - 0.0：与所有参考回复事实不符。
- **writing-style**：基于专业性、清晰度和简洁度评分。
```

**User Prompt:**

```
### 患者案例
标题：{title}
内容：{content}

### RAG CONTEXT
{rag_ctx}

### EXPERT REFERENCE (GOLD STANDARD)
{ref_text}

### CANDIDATE ANSWER TO EVALUATE
{candidate}

Provide your analysis and scores in structured JSON. You MUST provide a score for ALL metrics defined in the schema.
```