MedAware at MEDIQA-EVAL 2026: Vision-Language Model Fine-Tuning with Logprob-Based Score Calibration for Medical Response Evaluation

Ziqi Hao[1,*], **Pengbo Liu**[2,*]
[1]McGill University, [2]Unaffiliated
[1]3801 Rue University, Montreal, QC H3A 2B4, Canada
[1]ziqi.hao@mail.mcgill.ca, [2]liupengbo.work@gmail.com
*Corresponding authors

Abstract

We present MedAware, our MEDIQA-EVAL 2026 system for predicting human ratings of medical QA responses from text and images. We fine-tune Qwen3-VL models (4B/8B/32B) with supervised fine-tuning (SFT), and study GRPO as an optional second stage under both LoRA and full-parameter settings. To handle severe label skew and unstable correlation metrics, we use logprob-based continuous scoring with quantile calibration, converting token probabilities into calibrated metric scores without retraining. This reduces prediction collapse on skewed dimensions and improves metric stability in both English and Chinese. The approach follows the official reference-based shared-task setup and is designed to produce meaningful metric estimates even under extreme class imbalance. In the official shared-task submission setting (8B-LoRA SFT with discrete scoring), our system ranked 3rd on English and 1st among participants on Chinese. Separately, in post-competition offline re-evaluations with logprob scoring, the best tested configuration reaches 0.449 EN-ALL and 0.308 ZH-ALL, while SFT initialization remains critical for effective GRPO.

Keywords: medical evaluation, vision-language model, fine-tuning, score calibration, MEDIQA

1. Introduction

Automated evaluation of medical question answering (QA) systems is essential for scaling quality assessment beyond manual expert review. Unlike general-domain text evaluation, medical response assessment requires judging factual accuracy against clinical evidence, completeness of diagnostic reasoning, and appropriateness of language — often with associated clinical images. The MEDIQA-EVAL 2026 shared task (Ben Abacha and Yim, 2026; Yim et al., 2025a) formalizes this problem: given a medical case, a candidate system response, and gold-standard references, predict human expert ratings across multiple quality dimensions.

Recent work has demonstrated that large language models (LLMs) can serve as effective evaluators of text quality. Zheng et al. (2023) showed that GPT-4 judgments correlate strongly with human preferences on open-ended generation, and Liu et al. (2023) improved evaluation quality by leveraging chain-of-thought reasoning and output token probabilities. However, directly applying these LLM-as-judge approaches to structured medical evaluation presents three challenges: (1) the evaluator must understand domain-specific clinical content and integrate multimodal information (images, text); (2) the task requires simultaneous prediction across multiple quality dimensions with heterogeneous scales; and (3) several evaluation metrics exhibit highly skewed ground-truth distributions — for example, 96.7% of Chinese `writing-style` ratings are 1.0 — causing models to collapse to constant predictions with undefined correlation.

As team MedAware, we address these challenges through fine-tuned Qwen3-VL (Bai et al., 2025) vision-language models with a post-hoc calibration strategy. Our contributions are:

- A two-stage training formulation for multimodal medical evaluation, where supervised fine-tuning (SFT) provides a base evaluator and reinforcement learning (GRPO) is explored as an optional second-stage alignment method.
- A stage-aware comparison across model scales (4B, 8B, 32B): SFT-only, SFT→GRPO, and GRPO-only, while treating parameterization (LoRA vs. full fine-tuning) as an orthogonal choice within each stage. We report primary results on the official test set and use cross-validation for model selection.
- A logprob-based continuous scoring method with quantile calibration that addresses skewed label distributions and undefined correlations, yielding more stable and informative metric estimates.
- Empirical GRPO findings that SFT initialization is critical for GRPO to approach SFT performance and for logprob scoring to transfer effectively.

Proceedings of the 8th Workshop on Clinical Natural Language Processing (Clinical NLP) @ LREC 2026, pages 192–199

2. Related Work

LLM-based evaluation. The paradigm of using LLMs as evaluators has gained significant traction. Zheng et al. (2023) proposed using GPT-4 as a judge for open-ended generation, showing strong correlation with human preferences. Liu et al. (2023) introduced G-Eval, which leverages chain-of-thought reasoning and token probabilities to improve evaluation quality. Our work extends this line by fine-tuning a vision-language model specifically for medical evaluation with structured multi-metric prediction.

Medical NLP shared tasks. The MEDIQA series of shared tasks has driven progress in medical NLP evaluation, including multilingual and multimodal settings.

Reinforcement learning for alignment. Following DeepSeek-R1 (DeepSeek-AI, 2025), GRPO optimizes language models using group-relative advantages without requiring a separate value model. In medical vision-language reasoning, Med-R1 (Chen et al., 2025) further shows GRPO-style training can improve cross-modality and cross-task generalization. For medical image grounding, MedGround-R1 (Xu et al., 2025) adapts GRPO with spatial-semantic rewards and reports strong gains without chain-of-thought supervision. We explore GRPO for aligning model outputs with human evaluation judgments in medical response evaluation.

3. Task Description

The MEDIQA-EVAL 2026 task (Ben Abacha and Yim, 2026; Yim et al., 2025a) requires participants to predict human evaluation ratings for medical question answering systems. Given a medical case (clinical question and context, optionally with images), a candidate system response, and gold-standard reference responses, the system must predict quality ratings on a 3-point scale ($\{0, 0.5, 1.0\}$) across multiple dimensions.

English metrics (6). Each candidate response is rated on a 3-point scale $\{0, 0.5, 1.0\}$ for six dimensions: `disagree_flag` (factual contradiction/error; binary with 1 = disagree and 0 = agree); `completeness` (coverage of question aspects; 1 = complete, 0.5 = partial, 0 = incomplete); `factual-accuracy` (factual correctness against gold references); `relevance` (alignment with the patient question); `writing-style` (appropriateness for patient-facing communication); and `overall` (holistic quality across the above dimensions).

Chinese metrics (2). Chinese cases are rated on: `factual-consistency-wgold` (factual agreement with gold references) and `writing-style` (communication appropriateness). Note that the Chinese `writing-style` distribution is extremely skewed: 96.7% of ratings are 1.0, making correlation-based evaluation particularly challenging.

Evaluation. System performance is measured by the average of Kendall's τ, Pearson's r, and Spearman's ρ correlations between predicted and human ratings, computed per (language, metric) pair and then macro-averaged across all 8 metrics.

Data. The training set contains 161 medical cases (56 IIYI + 105 WoundCare) with 4,872 individual ratings. The IIYI subset is based on DermaVQA (Yim et al., 2024), and the WoundCare subset is based on WoundcareVQA (Yim et al., 2025b). The test set contains 193 cases with ratings from 4 human raters. Both English and Chinese versions are provided for all cases; representative bilingual examples are shown in Figure 1.

Following the official MEDIQA-EVAL 2026 setup, gold-standard reference responses are provided as part of the input at both training and test time. Our method is therefore a *reference-based* evaluator: at inference, the model conditions on the case, candidate response, and provided references, but never on the hidden human ratings used for evaluation.

4. System Description

Our full pipeline (Figure 2) is: model tuning (SFT, GRPO-only, or SFT→GRPO) produces a scorer that generates discrete JSON scores, and post-hoc logprob extraction converts these into continuous scores for calibration.

4.1. Problem Formulation

We formulate evaluation as structured JSON prediction. Given input context x comprising the clinical case (question, context, up to 3 images), a candidate system response, and up to 3 gold-standard reference responses, the model generates a single JSON object containing scores for all relevant metrics in one decoding pass. For English:

```
{"disagree_flag": 0, "completeness": 0.5, "factual-accuracy": 1.0, "relevance": 1.0, "writing-style": 1.0, "overall": 0.5}
```

This joint formulation captures inter-metric dependencies (e.g., low factual accuracy should imply low overall score) without requiring separate inference passes.

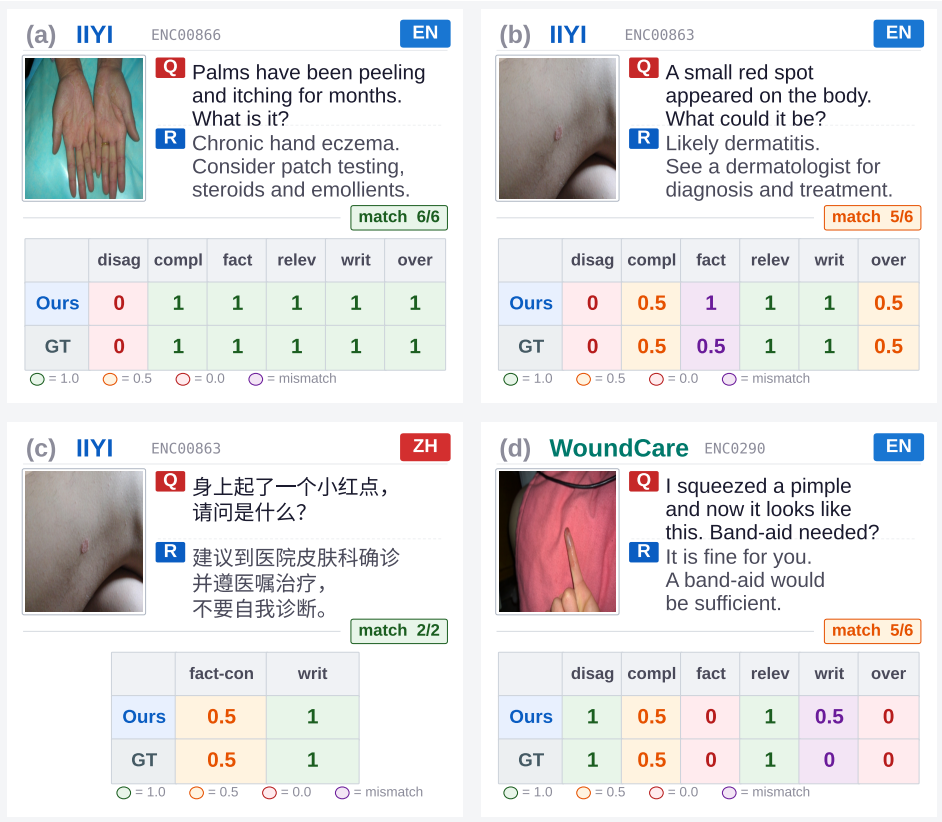

Figure 1: Representative MEDIQA-EVAL 2026 examples from IIYI/WoundCare in English and Chinese. Color badges denote human scores: 1.0, 0.5, 0.0.

Stage 1: Model Tuning (SFT / GRPO)
Input: case + images + candidate + gold refs
Backbone: Qwen3-VL + (LoRA or full FT)
Training mode: SFT / GRPO-only / SFT→GRPO
Output: `{"factual-accuracy": 1.0, "completeness": 0.5, ...}`
Stage 2: Logprob Extraction
At each score token position:
$\hat{s} = P(\text{“1”}) \cdot 1.0 + P(\text{“0”}) \cdot (\pi_{0.5} \cdot 0.5)$
Stage 3: Quantile Calibration
$\hat{s}_{\text{cal}} = F_{\text{GT}}^{-1}(\text{rank}(\hat{s}))$

Figure 2: Pipeline overview of our approach.

4.2. Model and Prompt Design

We use Qwen3-VL (Bai et al., 2025), a vision-language model supporting interleaved image-text inputs via a unified transformer architecture. We experiment with 4B, 8B, and 32B parameter variants.

The input prompt is structured as a chat-style conversation. The **system message** instructs the model to act as a medical expert and provides scoring guidelines (e.g., “Give 0.0 if the system response contradicts the gold standard”). The **user message** concatenates: (1) clinical images (encoded as vision tokens); (2) the patient question; (3) gold-standard reference responses; (4) the candidate response to evaluate; and (5) metric-specific scoring rubrics with an output format template. Separate prompt templates are used for English and Chinese, with language-appropriate metric definitions.

4.3. Supervised Fine-Tuning

We fine-tune using LoRA (Hu et al., 2022) with rank r=16, scaling factor α=32, and dropout 0.05, applied to attention and MLP projection layers. This yields approximately 44M trainable parameters for the 8B variant ($\sim$0.5% of total). We also compare against full fine-tuning for the 4B and 8B models.

Training minimizes cross-entropy loss on the target JSON tokens only, with all input prompt tokens masked (label $= -100$):

$$\mathcal{L}_{\text{SFT}} = -\sum_{t \in \mathcal{T}_{\text{output}}} \log P(y_t \mid y_{<t}, x; \theta) \qquad (1)$$

where $\mathcal{T}_{\text{output}}$ denotes token positions in the target JSON response.

Key hyperparameters: learning rate 2×10^{-5} (linear warmup over 5% of steps), per-device batch size 1 with gradient accumulation over 8 steps (effective batch size 8), AdamW with weight decay 0.01, gradient clipping at 1.0, bfloat16 preci-

sion. We train for 15–30 epochs depending on configuration (30 for 4B/8B LoRA, 20 for 32B-LoRA and 4B-Full, 15 for 8B-Full), evaluating on the held-out fold at every epoch. We select the checkpoint with the lowest validation loss via `load_best_model_at_end`; in practice, the best checkpoints consistently occur at epoch 2–4, with all models overfitting substantially thereafter due to the small training set (130 cases per fold), as shown in Figure 3.

4.4. Group Relative Policy Optimization

We explore GRPO, as applied in DeepSeek-R1 (DeepSeek-AI, 2025), to further align model outputs with human evaluation judgments as a post-competition experiment. We investigate two initialization strategies: (1) starting from an SFT checkpoint (**SFT→GRPO**), and (2) training directly from the pretrained base model (**GRPO only**).

We also run two reward families in exploratory GRPO experiments. The *simple* reward focuses on single-metric accuracy, providing a cleaner gradient signal that is easier to debug and interpret. The *composite* reward combines multi-metric accuracy with small bonuses for exact matches, output format, reasoning tags, and calibration:

$$r_{\text{simple}} = 0.2 \cdot r_{\text{fmt}} + 0.8 \cdot r_{\text{acc-single}}$$

$$r_{\text{composite}} = r_{\text{acc-multi}} + r_{\text{em}} + r_{\text{fmt}} + r_{\text{think}} + r_{\text{cal}}$$

where $r_{\text{fmt}} \in \{0, 1\}$ indicates valid JSON output, $r_{\text{acc-single}} = 1 - |\hat{s} - s^*|$ is distance-based accuracy on one focus metric, and $r_{\text{acc-multi}}$ is the weighted mean of distance-based per-metric accuracies. In the composite reward, $r_{\text{em}}{=}0.5\rho_{\text{exact}}$ uses the per-metric exact-match ratio $\rho_{\text{exact}} \in [0, 1]$, while $r_{\text{fmt}}{=}0.05b_{\text{fmt}}$, $r_{\text{think}}{=}0.05b_{\text{think}}$, and $r_{\text{cal}}{=}0.05b_{\text{cal}}$ depend on valid JSON, the presence of a `<think>...</think>` block, and use of valid score values. All GRPO experiments reported in this paper use thinking mode: the model first produces a brief `<think>...</think>` analysis and then outputs the final JSON. We compare these two reward families as post-competition diagnostics; rollout count K is not treated as an ablation variable in this paper.

GRPO computes per-sample advantages normalized within each group of K candidates:

$$a_j = \frac{r_j - \bar{r}}{\max(\sigma_r, \epsilon)} \tag{2}$$

and updates the policy via a clipped surrogate objective with KL regularization ($\beta{=}0.1$) against the frozen reference model. We use conservative hyperparameters: learning rate $5{\times}10^{-7}$ and max generation length 768 tokens, and train up to 3000 episodes. Unless otherwise stated, all GRPO inference results reported in this paper use checkpoint-1000. At inference time, we apply the same logprob-based scoring and quantile calibration pipeline described in Sections 4.5–4.6.

4.5. Logprob-Based Continuous Scoring

A key challenge is that several metrics have highly skewed ground-truth distributions. For example, Chinese `writing-style` has 96.7% of ratings equal to 1.0. Models trained on discrete targets $\{0, 0.5, 1.0\}$ learn to predict the majority class for all samples, producing zero-variance predictions where correlation is undefined.

We address this by extracting continuous scores from the model's output token probabilities at inference time, requiring no retraining. We run generation with `output_scores=True` to obtain per-position logit vectors. For each metric value in the generated JSON, we locate the corresponding score token position via character-to-token alignment, then compute:

$$\hat{s} = \tilde{P}_1 \cdot 1.0 + \tilde{P}_0 \cdot (\pi_{0.5} \cdot 0.5) \tag{3}$$

where $\tilde{P}_0, \tilde{P}_1$ are the softmax probabilities of tokens "0" and "1" renormalized to sum to 1, and $\pi_{0.5} = P(\text{score}{=}0.5 \mid \text{score}{\neq}1.0)$ is the prior probability of a 0.5 rating among non-1.0 ratings, estimated per metric from the training set (e.g., $\pi_{0.5}{=}0.738$ for `completeness`, 0.0 for binary `disagree_flag`).

We compute $\hat{s}$ using the first digit token ("0" vs. "1") rather than directly using $P(\text{"0.5"})$ because (i) "0.5" is often tokenized into multiple sub-tokens, making probability extraction sensitive to exact token boundaries, and (ii) the first digit token can be aligned more reliably across decoding variants and JSON formatting differences. The class prior $\pi_{0.5}$ then splits the "0" mass between 0.0 and 0.5 in expectation.

The intuition is that even when the model's argmax output is "1" for every sample, the probability mass on "0" varies and captures the model's relative uncertainty. This converts a degenerate constant prediction into a continuous signal with nonzero variance.

4.6. Quantile Calibration

Logprob scores preserve ranking but may not match the ground-truth scale, which matters for Pearson correlation. This is consistent with the broader calibration literature showing that modern neural networks can be miscalibrated (Guo et al., 2017). We apply per-metric quantile calibration: each prediction is mapped to its percentile rank among all predictions for that metric, then mapped to the corresponding quantile of the training set's

ground-truth distribution:

$$\hat{s}_{\text{cal}} = F_{\text{GT}}^{-1}(\text{rank}(\hat{s})) \quad (4)$$

where F_{GT}^{-1} is the empirical quantile function of human ratings. This aligns the marginal distribution of predictions with human ratings while preserving the prediction ranking.

Importantly, calibration uses only training-set labels to estimate F_{GT}; on validation/test splits it uses only the model predictions to compute $\text{rank}(\hat{s})$ (no access to evaluation labels).

5. Experimental Setup

5.1. Cross-Validation

We perform 5-fold cross-validation on the training set (161 cases) by splitting at the encounter/case level and stratifying folds by dataset (IIYI vs. WoundCare) to ensure balanced representation. Each fold trains on 80% of cases and evaluates on the held-out 20%. This allows us to compare model configurations without touching the test set.

5.2. Configurations

We evaluate five configurations: **4B-LoRA**, **8B-LoRA**, and **32B-LoRA** (all with LoRA rank 16), plus **4B-Full** and **8B-Full** (full fine-tuning). Training uses bfloat16 on NVIDIA H100 80GB GPUs, with 1–4 GPUs depending on configuration and training stage.

6. Results

6.1. Cross-Validation Results

As shown in Table 1, 32B-LoRA attains the highest EN (0.481) and ZH (0.374) under this CV protocol, and is highest on each English metric in this table. Among metrics, `factual-accuracy` is the easiest (∼0.49–0.52) while `writing-style` is the hardest (∼0.31–0.39) for English. Chinese `writing-style` is extremely difficult (∼0.14–0.17) due to distributional skew (96.7% of ground truth ratings are 1.0). LoRA and full fine-tuning perform comparably in CV: full fine-tuning is slightly better at 4B (EN 0.439 vs. 0.435) but worse at 8B on ZH (0.347 vs. 0.360), suggesting that neither consistently dominates on this small dataset.

Compared with discrete scoring, logprob-based scoring mitigates prediction collapse on highly skewed metrics (e.g., Chinese `writing-style`), yielding a continuous signal with nonzero variance and thus well-defined correlation estimates. Quantile calibration further aligns the prediction scale with the training-set rating distribution while preserving ranking.

6.2. Test Set Results

Table 2 reports post-competition offline test-set logprob results for the five SFT configurations and the GRPO models. Table 2 includes only composite-reward GRPO rows; simple-reward 8B diagnostics appear in Table 3. Our official submission was 8B-LoRA with discrete scoring, ranking 3rd on English and 1st among participants on Chinese. These offline test-set numbers should therefore be interpreted separately from the official leaderboard results. Model rankings shift between CV and test: 32B-LoRA leads CV (EN 0.481), while 4B-Full leads test (EN 0.449), consistent with high variance under the small training set (161 cases).

6.3. GRPO Results (Post-Competition)

As shown in Table 2, SFT→GRPO (LoRA) remains close to SFT performance with logprob scoring: EN is 0.428 (4B) and 0.431 (8B), versus 0.437 and 0.433 for the corresponding SFT-LoRA models. Per-metric behavior is similar, with `factual-accuracy` strongest (0.542–0.544) and `completeness` weakest (0.309–0.322). In contrast, GRPO-only settings (LoRA and Full FT) underperform their matched SFT baselines across scales, with EN drops of 0.072–0.111 and `overall` drops of 0.071–0.127.

6.4. Additional GRPO Diagnostics

Table 3 shows that initialization matters more than reward family: within each reward family, SFT→GRPO clearly outperforms GRPO only (EN-logprob gain +0.073–0.074). Figure 4 plots the first 1000 training steps for direct comparison.

LoRA vs. full fine-tuning under GRPO. With composite-reward logprob scoring, GRPO-only LoRA is slightly better on EN at both scales and on ZH at 4B, while Full FT is slightly better on ZH at 8B.

Logprob vs. discrete scoring. Table 4 compares both scoring methods for all five SFT models. Logprob scoring consistently improves EN-ALL, with the largest gains for full fine-tuned models.

For Chinese evaluation, discrete scoring yields undefined ZH-ALL for 3 of 5 SFT models because `writing-style` collapses to a constant prediction. Logprob scoring resolves this and improves EN for all five SFT models; gains are smaller for GRPO (+0.015–0.034 EN). This makes logprob scoring not merely a performance improvement but a practical requirement for evaluating highly skewed metrics with meaningful correlation estimates.

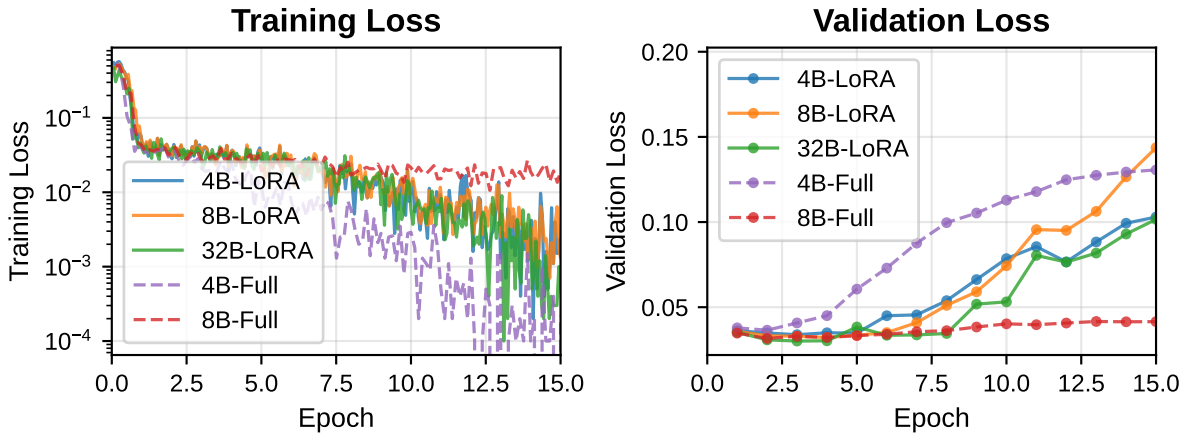

Figure 3: Training and validation loss curves for fold-0 across all configurations. All models overfit after epoch 5–10, with full fine-tuning (dashed) overfitting more severely than LoRA (solid). 32B-LoRA shows the slowest validation loss increase.

	English Metrics							Chinese		
Model	Disag.	Compl.	Fact.	Relev.	Writ.	Over.	**EN**	Fact.	Writ.	**ZH**
4B-LoRA	.429	.406	.490	.463	.333	.486	.435	.565	.169	.367
8B-LoRA	.411	.380	.485	.436	.310	.489	.418	.571	.149	.360
32B-LoRA	**.492**	**.422**	**.517**	**.524**	**.391**	**.542**	**.481**	.578	**.170**	**.374**
4B-Full	.410	.419	.467	.488	.380	.469	.439	**.585**	.156	.370
8B-Full	.387	.389	.499	.450	.355	.474	.426	.558	.137	.347

Table 1: 5-fold CV results with logprob scoring. Each cell is the mean across folds of the averaged (τ, r, ρ) score. EN = average of 6 English metrics; ZH = average of 2 Chinese metrics. Disag. = disagree_flag, Compl. = completeness, Fact. = factual-accuracy/consistency, Relev. = relevance, Writ. = writing-style, Over. = overall.

	English Metrics							Chinese		
Model	Disag.	Compl.	Fact.	Relev.	Writ.	Over.	**EN**	Fact.	Writ.	**ZH**
SFT (trained on CV fold-0, 80% of data):										
4B-LoRA	.385	.336	**.547**	.487	.352	.518	.437	.458	.034	.246
8B-LoRA†	.391	.321	.540	.461	.359	.524	.433	.484	.038	.261
32B-LoRA	**.430**	**.362**	.524	.461	.374	.506	.443	.475	**.100**	.288
4B-Full	.392	.338	.532	**.514**	**.412**	.508	**.449**	**.527**	.088	**.308**
8B-Full	.384	.297	.533	.440	.361	.502	.420	.482	.040	.261
SFT→GRPO (LoRA):										
4B	.387	.322	.542	.472	.337	.510	.428	.457	.043	.250
8B	.399	.309	.544	.465	.336	**.533**	.431	.463	.047	.255
GRPO only (LoRA):										
4B	.345	.239	.444	.371	.260	.406	.344	.323	.063	.193
8B	.376	.227	.509	.340	.244	.444	.357	.383	.025	.204
GRPO only (Full FT):										
4B	.345	.239	.431	.386	.247	.381	.338	.313	.059	.186
8B	.376	.215	.502	.319	.245	.431	.348	.386	.034	.210

Table 2: Test set results (193 cases) with logprob scoring. SFT models use CV fold-0 training data; GRPO rows use composite reward with $K{=}8$. † marks the official submission model; the numbers shown are post-hoc logprob re-evaluations. Red/blue denote best/second-best per column.

6.5. Analysis

Per-metric patterns. `factual-accuracy` is the strongest EN metric across all models (0.43–

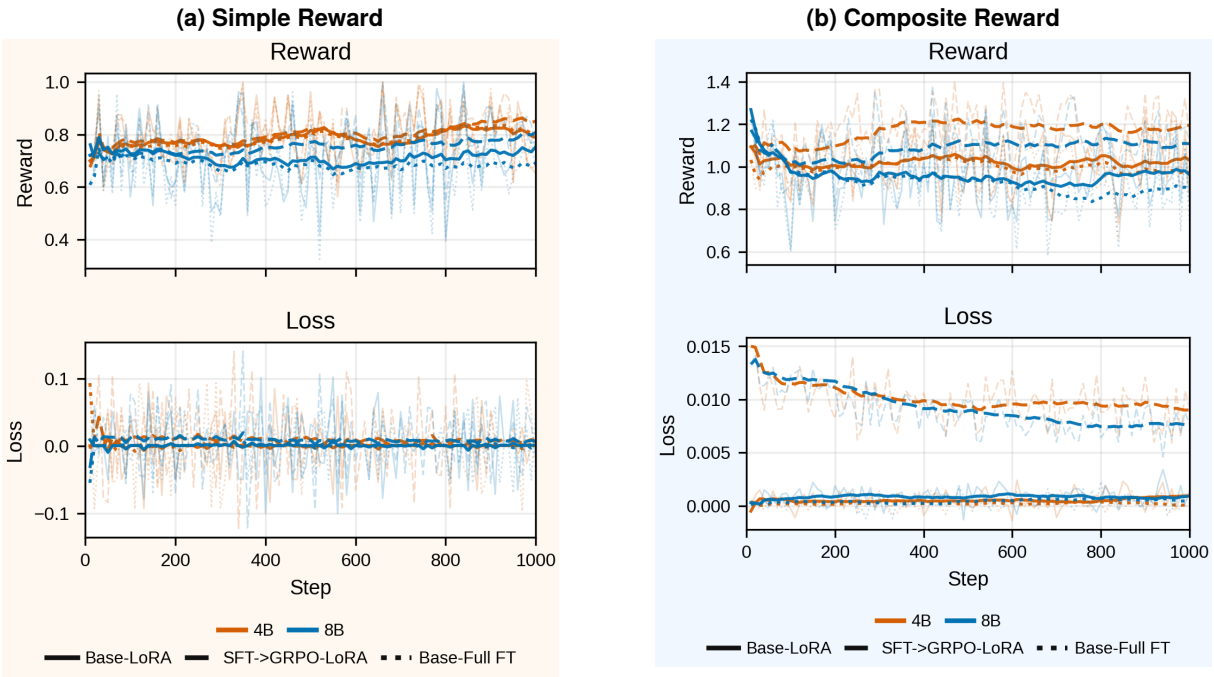

Figure 4: GRPO training curves over the first 1000 steps for simple and composite rewards. Colors denote model scale; line styles denote training strategy (GRPO only LoRA, SFT→GRPO LoRA, GRPO only Full FT).

		Discrete		Logprob	
Reward	**Init**	**EN**	**ZH**	**EN**	**ZH**
Simple	GRPO only	.338	.218	.355	.206
Simple	SFT→GRPO	**.405**	.231	.428	.252
Composite	GRPO only	.341	.218	.357	.204
Composite	SFT→GRPO	.397	**.236**	**.431**	**.255**

Table 3: 8B LoRA GRPO diagnostics comparing reward families and initialization strategies. The composite SFT→GRPO logprob row is the same run as the 8B SFT→GRPO row in Table 2.

	EN			ZH	
Model	Disc.	Logp.	Δ	Disc.	Logp.
4B-LoRA	.399	.437	+.038	—*	.246
8B-LoRA†	.404	.433	+.029	.254	.261
32B-LoRA	.395	.443	+.048	.308	.288
4B-Full	.370	.449	+.079	—*	.308
8B-Full	.296	.420	+.124	—*	.261

Table 4: Discrete vs. logprob scoring on the test set (SFT models). The Logp. columns correspond to the EN/ZH columns in Table 2. †Official submission. *ZH-ALL is undefined because discrete predictions for Chinese `writing-style` collapse to a constant value.

0.55), consistent with CV findings. `completeness` is consistently challenging (0.22–0.36). Chinese `writing-style` is near zero on test (0.03–0.10), far below CV values (0.14–0.17), suggesting poor generalization for this extremely skewed metric.

Model scale and configuration. Under this test split, 4B-Full is the best SFT model (EN 0.449, ZH 0.308), while 32B-LoRA is the strongest LoRA model (EN 0.443). Full fine-tuning also preserves richer logit distributions: 4B-Full is higher than 4B-LoRA on `relevance` (0.514 vs. 0.487) and `writing-style` (0.412 vs. 0.352) with logprob scoring.

CV-to-test gap. The ranking shift from 32B-LoRA in CV to 4B-Full on test suggests substantial fold variance with only 161 cases, making CV unreliable for separating closely-performing configurations. It also suggests that with limited medical evaluation data, smaller fully fine-tuned models may generalize more robustly than larger LoRA-adapted ones.

GRPO analysis. SFT→GRPO preserves the logit structure needed for effective logprob scoring but does not improve on SFT. The likely causes are the small dataset and the low-entropy score space $\{0, 0.5, 1.0\}$, which often collapses reward variance within a GRPO group. GRPO only fails more severely because the pretrained model lacks the structured JSON behavior that both the reward function and logprob extraction rely on.

7. Conclusion

We presented MedAware, a Qwen3-VL-based system for medical response evaluation. Our official submission used 8B-LoRA SFT, ranking 3rd on English and 1st among participants on Chinese. The main finding is that logprob-based continuous scoring with quantile calibration is essential under skewed rating distributions, turning unstable or undefined discrete correlations into usable evaluation signals. For GRPO, the clearest result is the gap between initialization strategies: SFT→GRPO remains competitive, whereas GRPO-only substantially underperforms, indicating that supervised structured generation is a prerequisite for effective RL in this setting.

8. Limitations

Our approach has several limitations. The dataset is small (161 cases), which limits generalization and makes correlation-based comparisons unstable. The evaluator is reference-based, which matches the shared-task setup but limits direct use in settings without expert reference responses. Logprob scoring relies on priors estimated from the training distribution and may transfer poorly under distribution shift. GRPO was explored only in a limited regime, and longer training, larger-scale runs, and stronger cross-metric calibration remain future work.

9. Acknowledgements

We thank the MEDIQA-EVAL 2026 organizers for creating this shared task and providing the dataset.

10. Bibliographical References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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*. Association for Computational Linguistics.

Haotian Chen, Han Ding, Yiru Bai, Ren Gao, Junbo Yang, Yuheng Li, and Konstantinos Psounis. 2025. Med-R1: Reinforcing medical reasoning in MLLMs via GRPO with thought preferences. *arXiv preprint arXiv:2503.13939*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Edward J. Hu, Yelong Shen, Phillip Wallis, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Huihui Xu, Yuanpeng Nie, Hualiang Wang, et al. 2025. MedGround-R1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15964, pages 391–401. Springer Nature Switzerland.

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. *CoRR*, abs/2509.12405.

Wen-Wai Yim, Asma Ben Abacha, Robert Doerning, et al. 2025b. Woundcarevqa: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888.

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 *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005, pages 209–219. Springer Nature Switzerland.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.