CARES at MediaEval Medico 2026: Calibrated Aspect-Reliable Explanations and Safety for Gastrointestinal Visual Question Answering

Syed Saad Hasan Emad[1,*], Itbaan Safwan[1] and Muhammad Atif Tahir[1]

[1]*School of Mathematics and Computer Science, Institute of Business Administration (IBA), Karachi, Pakistan*

Abstract

We describe *CARES*, our submission to the Medico 2026 task on visual question answering for gastrointestinal endoscopy [1], in both subtasks of the Kvasir-VQA-x1 [2] benchmark. CARES is a training-free reliability layer wrapped around a frozen Florence-2-base model with a public LoRA adapter; we did not fine-tune the backbone. The layer contributes (i) a per-(aspect, answer) reliability `confidence_score` that reaches AUROC 0.87, well above max-softmax (∼0.60) and self-consistency (∼0.62); (ii) a self-probing, faithful-by-construction explanation template; (iii) a distribution-free conformal selective-risk control that, with exact Clopper–Pearson bounds at $\delta = 0.1$, gives 55% coverage at a verified $\leq$5% selective error in every fold we tried; and (iv) a triangulated faithfulness evaluation showing the template dominates LLM narration on three independent metrics. The unifying empirical finding is *base-rate dominance*: on a frozen medical VLM, correctness is governed by output-conditioned base rates, not by any image-conditional or generative signal we tested.

1. Introduction

The Medico 2026 task [1] asks for accurate but also *trustworthy* answers on gastrointestinal endoscopy images: explanations should be faithful and visually grounded, and answers should carry honest uncertainty so wrong predictions can be safely set aside. We refer to the overview paper for the task definition and dataset details, and to the Kvasir-VQA-x1 paper [2] for the underlying benchmark.

Our entry, CARES, asks how far a frozen, modestly-sized vision–language model can be pushed toward this kind of trustworthiness by an inference-time layer alone. Two design properties of the data drove this choice. First, the verbose reference answers on Kvasir-VQA-x1 are not classification labels: they are short clinical sentences with synonym variation ("evidence of colonoscopy procedure" vs. "evidence of colonoscopic examination"), which makes BLEU on the deployed backbone close to its paraphrasing-bounded ceiling, but leaves ample headroom on the safety and explanation side. Second, intrinsic model-confidence signals were uninformative on this task in our preliminary experiments, which pushed us toward output-side calibration rather than instance-side uncertainty. Section 3 makes the latter point quantitative: it underlies what we call *base-rate dominance* and is the most transferable insight from our submission.

The closest prior work is the MediaEval Medico 2025 entry of Safwan et al. [4], which fine-tunes Florence-2 with LoRA for multi-task VQA, explanation, and grounding. We

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
EMAIL: syed.saad.31916@khi.iba.edu.pk (S. S. H. Emad); atiftahir@iba.edu.pk (M. A. Tahir)

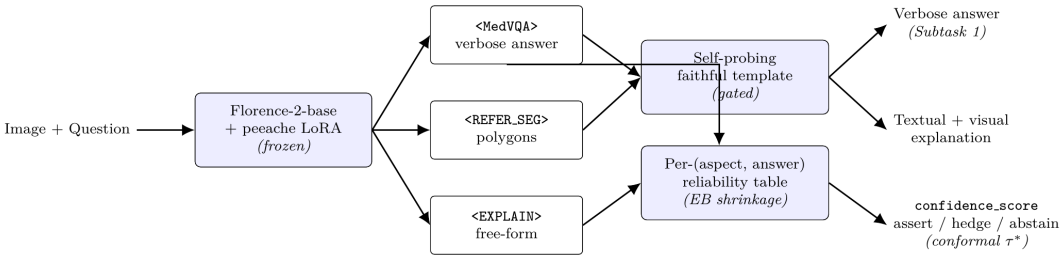

Figure 1: CARES pipeline. The frozen Florence-2 backbone emits the verbose answer, a referring segmentation, and a free-form explanation. The self-probing template anchors a clinician-style narrative on the validated answer and overlay; a per-(aspect, answer) reliability table turns the answer into a calibrated probability that drives the safety hedge.

keep the backbone frozen and put the effort into the reliability layer.

2. Approach

Backbone. Both subtasks use `microsoft/Florence-2-base` [3] with the public LoRA adapter `peeache/FL2_VQA_MIX_128_256`, frozen end-to-end. Beam search (N=3) and the official post-processors are used throughout. We run inference per-example: Florence-2 merges image tokens into its encoder sequence, and batched padding leaves the attention mask the wrong shape (a known SDPA mismatch).

Subtask 1: verbose answer generation. We feed the prompt `"<MedVQA> {q}"` and emit the model's free-form output directly. The reference answers in Kvasir-VQA-x1 are verbose, so no concise normaliser is applied; in pilot runs, mapping to a small controlled vocabulary collapsed BLEU from ∼0.48 to ∼0.004 because verbose references no longer matched the short normalised predictions.

Subtask 2: explanation and safety layer. We add three artefacts per case (Figure 1). **(B) Visual grounding.** When the answer is localizable (contains `polyp, ulcer, oesophag, snare, forceps, instrument, cecum, z-line, lesion, paris`), we issue `<REFERRING_EXPRESSION_SEGMENTATION>`, decode the ⟨`loc_x`⟩ tokens into polygons, and render a translucent overlay. A per-image grounding step propagates entity-typed masks to sibling questions on the same image without crossing entity types. **(C) Textual explanation.** We anchor the explanation on the validated answer (*"A polyp measuring approximately 5–10 mm is identified."*) and add tentative clauses built from the model's own attribute answers on sibling questions (location, colour, size, morphology), gated to keep absent findings, instruments and landmarks from picking up lesion attributes. The free-form `<MedVQA_EXPLAIN>` narrative is appended only if it does not contradict the validated answer and adds >50% novel non-stop content words. A verbal confidence clause closes the text. **(D) Calibrated confidence.** The `confidence_score` is a per-(aspect, answer) reliability $\hat{r}(a, \hat{y}) = (c + \alpha\bar{r}(a))/(n + \alpha)$ with $\alpha = 5$, fit on an image-disjoint, held-out calibration slice with empirical-Bayes shrinkage toward the per-aspect mean. The deployed safety policy is heuristic (assert at ≥0.80, hedge 0.50–0.80, abstain <0.50); Section 3 reports the distribution-free conformal upgrade.

3. Results and Analysis

3.1. Subtask 1 scores

Table 1 gives the official public scores produced by `submission_task1.py` on the Medico 2026 validation set (the first 1,500 items of the complexity-1 Kvasir-VQA-x1 test partition after a fixed shuffle).

Table 1
Subtask 1 public scores on the Medico 2026 validation set (1,500 verbose-answer items).

Model	BLEU	R-1	R-2	R-L	METEOR
CARES (ours)	0.4835	0.7162	0.5352	0.6895	0.6898

3.2. Subtask 2 internal diagnostics

Calibration. Table 2 compares confidence signals on the validation set with the reliability table fit out-of-fold (5-fold split on `img_id`). The per-(aspect, answer) score reaches AUROC 0.87 and Expected Calibration Error 0.025; every instance-level signal we tried is at or below 0.62.

Table 2
Confidence signals on the validation set (5-fold cross-fitted by `img_id`).

Signal	AUROC	ECE
Max-softmax probability (MSP)	0.60	0.06
MSP + temperature scaling	0.60	0.03
Explanation-generation likelihood	0.59	–
Answer self-consistency ($T = 5$)	0.62	–
Visual retrieval similarity (top-k FAISS)	0.45	–
Segmentation-grounding (area + has-mask)	0.51	–
Per-aspect reliability	0.77	0.04
Per-(aspect, answer) reliability	**0.87**	**0.025**

Selective prediction with a distribution-free guarantee. Choosing the smallest threshold τ^* for which the exact Clopper–Pearson upper bound on the selective risk satisfies $\widehat{R}^{+}(\tau^*) \leq \alpha$ at confidence $1 - \delta$, on a 3-way disjoint split (scorer-fit / pick-τ / verify), we get $\alpha = 0.05$ at 0.559 ± 0.051 coverage with 0.030 ± 0.006 verified selective error, $\alpha = 0.10$ at 0.822 ± 0.030 coverage with 0.077 ± 0.009 error, and $\alpha = 0.15$ at 0.938 ± 0.009 coverage with 0.122 ± 0.005 error. The guarantee held in every fold at every α. The deployed heuristic policy (0.80/0.50 thresholds) reaches asserted-answer accuracy 0.95, a confidently-wrong rate of 0.037, and recall of errors 0.76.

Explanation faithfulness. We measure faithfulness three ways on the same cases (Table 3): structured claim support against the case evidence; NLI entailment with `MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli`; and an LLM-as-judge using `Qwen2.5-3B-Instruct`. The deployed template wins on all three. The keyword-gated LLM rewrite barely recovers anything because semantic drift slips past a lexical gate: the gated and ungated variants are indistinguishable to the LLM-judge.

Table 3
Fluency–faithfulness frontier on Medico 2026 cases.

Variant	Structured	NLI ent.	LLM-judge
Template (deployed)	**0.951**	**0.736**	**0.812**
LLM narration, keyword-gated	0.884	0.551	0.784
LLM narration, ungated	0.862	0.538	0.792

3.3. Insights

The cleanest finding is that intrinsic and image-conditional signals (MSP, entropy, self-consistency, retrieval, segmentation grounding) all fall at or below AUROC 0.62 for predicting correctness on this frozen backbone, while a transparent per-(aspect, answer) base-rate table reaches 0.87. We refer to this as *base-rate dominance*. Practically, on a frozen medical VLM the question's aspect together with the value of the answer encodes most of the available information about whether the answer is right; the image-side signals add little. The faithfulness results sharpen the same lesson on the explanation side: a faithful-by-construction template beats a fluent LLM rewriter, and a lexical gate over LLM rewrites yields false reassurance because the LLM-judge cannot tell gated from ungated apart.

4. Discussion and Outlook

Base-rate dominance is the most transferable point from our submission. It suggests that calibration work on frozen medical foundation models should treat output-conditioned base rates as a strong baseline before reaching for instance-level signals, and that distribution-free risk control on top of such a base-rate score can yield clinically interpretable abstention with provable guarantees on small calibration sets. Three further attempts that did not help and informed the design: decoding tuning (within the ± 0.005 BLEU noise band), a vision-only LoRA continuation of the backbone (BLEU $+0.0008$, no help), and a 1-epoch language-model LoRA SFT on 143k pairs (BLEU -0.0005, METEOR $+0.007$ — not enough to redeploy). The unifying interpretation is again base-rate dominance: the remaining gap is paraphrasing-bounded for this backbone, and the intervention most likely to move it substantially is a change of backbone. Future work will therefore either swap to a stronger fine-tunable backbone for Subtask 1 or augment the reliability layer with a learned uncertainty head on frozen multimodal features that consults the same base-rate prior as an inductive bias.

Declaration on Generative AI

During the preparation of this work, the authors used a large-language-model assistant (*Grammar and spelling check*; *Paraphrase and reword*) and an LLM-based help for table and diagram refinement. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Gautam, V. Thambawita, M. A. Riegler, and others. Medico 2026: Visual Question Answering for Gastrointestinal Imaging. In *Proc. of the MediaEval 2026 Workshop*, Amsterdam, Netherlands and Online, 15–16 June 2026.

[2] S. Gautam, M. A. Riegler, and P. Halvorsen. Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy. arXiv:2506.09958, 2025.

[3] B. Xiao et al. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *CVPR*, 2024.

[4] I. Safwan, M. A. Shaikh, M. Haaris, R. Khan, M. A. Tahir. Multi-task Learning for Visually Grounded Reasoning in Gastrointestinal VQA. arXiv:2511.04384, 2025.

[5] D. Hendrycks and K. Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *ICLR*, 2017.

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *ICML*, 2017.

[7] Y. Geifman and R. El-Yaniv. Selective Classification for Deep Neural Networks. In *NeurIPS*, 2017.

[8] S. Bates, A. N. Angelopoulos, L. Lei, J. Malik, M. I. Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *JACM*, 2021.

[9] A. N. Angelopoulos and S. Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, 2023.

[10] Q. Lyu et al. Towards Faithful Model Explanation in NLP: A Survey, 2024.

[11] P. Atanasova et al. Faithfulness Tests for Natural Language Explanations. In *ACL*, 2023.