

CARES: Calibrated Aspect-Reliable Explanations and Safety for Gastrointestinal VQA

Syed Saad Hasan Emad¹, Itbaan Safwan¹ and Muhammad Atif Tahir¹

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

Abstract

This paper describes our submission to ImageCLEFmedical 2026 MEDVQA-GI. We took part in both subtasks: visual question answering on Kvasir-VQA-test (Subtask 1) and explainable, safety-aware reasoning (Subtask 2). CARES is a training-free reliability layer built around a frozen Florence-2-base model with the public peeache/FL2_VQA_MIX_128_256 LoRA adapter; we did not fine-tune the backbone for either subtask.

The reliability layer has four parts. First, a calibrated `confidence_score` obtained from a per-(aspect, answer) reliability table fit by empirical-Bayes shrinkage on an image-disjoint calibration slice. The score reaches AUROC 0.87 for predicting answer correctness, well above the model's own max-softmax probability (~ 0.60), temperature scaling, and sample self-consistency (~ 0.62). Second, a self-probing template that turns the model's own attribute answers (location, colour, size, morphology) into a tentative clinician-style report; the synthesis is gated so absent findings, instruments, and landmarks never receive lesion attributes (zero phantom cases on the validation set). Third, a safety policy keyed to the calibrated probability (assert at ≥ 0.80 , hedge between 0.50 and 0.80, abstain below 0.50) that flags about 76% of errors while keeping the confidently-wrong rate at 3.7%. Fourth, a set of honest negative results: every image-conditional signal we tried (MSP, self-consistency, visual retrieval over a FAISS train index, segmentation-grounding quality, LLM rephrasing of explanations) was either useless or actively harmful for predicting correctness.

We also evaluate explanation with three independent metrics (structured claim support, NLI entailment with DeBERTa, and an LLM-as-judge). The faithful-by-construction template wins on all three; a lexical faithfulness gate over LLM rewrites looks like it helps but does not, because semantic drift passes straight through.

The pattern across these results is what we call *base-rate dominance*: on a frozen medical VLM, correctness is governed mostly by output-conditioned base rates, not by any image-conditional or generative signal. A small, transparent reliability table is therefore both the safer and the more accurate choice.

1. Introduction

Clinical visual question answering systems need to be accurate, but they also need to be trustworthy: explanations should be faithful and visually grounded, and answers should be paired with honest uncertainty so that an incorrect prediction can be safely set aside. Rather than build a bigger model, this paper asks how far a frozen, modestly-sized vision-language model can be pushed toward this kind of trustworthiness by an inference-time layer alone. The whole pipeline fits on an 8 GB consumer GPU; keeping the backbone fixed isolates the contribution of the reliability layer from any improvement that might come from a stronger underlying model. The MEDVQA-GI 2026 challenge [1, 2] reflects a similar turn toward interpretability and safety, scoring Subtask 1 with BLEU/ROUGE/METEOR and Subtask 2 with a multimodal explanation and a behavioural safety rubric. Both subtasks are handled with one shared model and one shared reliability layer.

Our contributions are four. First, a calibrated per-(aspect, answer) reliability table fit by empirical-Bayes shrinkage on an image-disjoint slice; deployed as `confidence_score` it reaches AUROC 0.87, against 0.60 for max-softmax and 0.62 for sample self-consistency. Second, a self-probing explanation template whose synthesis is gated so absent findings, instruments, and landmarks never receive lesion attributes (zero phantom cases on validation). Third, a calibrated safety policy with a verified distribution-free selective-risk guarantee via an exact Clopper-Pearson upper bound [11]. Fourth, a set of honest negative results: max-softmax, self-consistency, visual retrieval over a FAISS train index,

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

EMAIL: syed.saad.31916@khi.iba.edu.pk (S. S. H. Emad); atiftahir@iba.edu.pk (M. A. Tahir)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and segmentation grounding all sit near AUROC 0.6 or below, while output-conditioned base rates dominate. The code and models are available at <https://github.com/saademad200/CARES-MedVQA>.

2. Method

Backbone. Both subtasks load `microsoft/Florence-2-base` with the public `peeache/FL2_VQA_MIX_128_256` LoRA adapter. The backbone is frozen end-to-end; only the calibration table and the conformal thresholds are data-dependent. Images are resized to 768×768 . Inference fits on an 8 GB RTX 4070 Laptop with greedy decoding.

Subtask 1: concise answer generation. The medical-VQA task token produces a verbose answer that is mapped to the controlled-vocabulary references of Kvasir-VQA-test by a deterministic concise-normaliser. The normaliser is the strongest single contributor to the public lexical scores. Table 1 details the per-aspect rewrite rules.

Aspect	Rule
artefact, text_presence color	Yes/No canonicalisation: any negative keyword (<i>no</i> , <i>none</i> , <i>not visible</i>) maps to “No”; otherwise “Yes”. Colour tokens extracted with a closed vocabulary of 12 hues; multiple colours sorted alphabetically and joined with semicolons to ensure exact-match with multi-label references (e.g. “pink;red;white”).
location	A 9-cell grid (upper/center/lower $\times$ left/center/right) extracted via regex on directional keywords; all detected cells sorted alphabetically and joined with semicolons.
polyp_count, instrument_count	Number words (<i>one</i> $\rightarrow$ 1, <i>two</i> $\rightarrow$ 2, ...) converted; first digit extracted; “no / none” maps to 0; cross-contamination blocked (instrument-count ignores polyp-only vocabulary).
polyp_type	Pedunculated (Paris Ip), sessile (Paris Is), and flat (Paris IIa/b/c) synonyms collapsed into the controlled Paris vocabulary.
polyps (size)	Size ranges (<5 mm, 5–10 mm, 11–20 mm, >20 mm) extracted; multi-range answers joined with semicolons.
polyp_removal	Any verbose phrase longer than three tokens is mapped to “not relevant” to avoid spurious partial matches.
procedure	Keyword match on <i>gastroscopy</i> / <i>colonoscopy</i> ; all other answers pass through unchanged.

Table 1: Per-aspect rewrite rules of the concise-normaliser.

Subtask 2: explanation and safety. The validated Subtask-1 answer is passed through the self-probing template. The template issues up to four follow-up queries to the same frozen model about the finding: location, colour, size, and morphology. A hard gate blocks lesion attributes whenever the initial answer is *none*, an instrument, or an anatomical landmark, eliminating all phantom lesion descriptions on the validation set. The responses are assembled into a tentative clinical sentence.

Self-probing example. Given image `cju0qkw191gd1` with question “What type of polyp is present?”:

1. Initial answer: *Paris Is (sessile)*.
2. Location probe: “Where in the image is the polyp?” $\rightarrow$ *lower-center*.
3. Colour probe: “What colour is the polyp?” $\rightarrow$ *pink*.

4. Size probe: “What is the size of the polyp?” $\rightarrow$ 5–10 mm.
5. Assembled: “A 5–10 mm, pink, sessile (Paris Is) polyp is located in the lower-center region. The referring-expression segmentation confirms the lesion boundary.”

The referring-expression segmentation token (`REFERRING_EXPRESSION_SEGMENTATION`) accepts the concise answer as its text query and returns Florence-2 location tokens (`loc_x`) encoding polygon vertices on a 1000×1000 grid. Pixel coordinates are recovered as $x = (X/1000) \cdot w, y = (Y/1000) \cdot h$ and rendered as a semi-transparent green overlay saved to the `visuals/` directory bundled with the JSONL submission.

A per-(aspect, answer) reliability table $\hat{r}(a, \hat{y}) = (c(a, \hat{y}) + \alpha \bar{r}(a)) / (n(a, \hat{y}) + \alpha)$ converts the answer to a calibrated probability. Here $c(a, \hat{y})$ is the empirical correct count, $n(a, \hat{y})$ the total count, and $\bar{r}(a)$ the aspect-level mean, all from an image-disjoint calibration slice. The shrinkage parameter $\alpha = 5$ was selected by grid search over $\{1, 5, 10, 20\}$ on the calibration slice to minimise Expected Calibration Error (ECE), yielding $\text{ECE} = 0.025$ (cross-fitted). Sensitivity analysis shows AUROC remains stable at 0.86–0.87 for $\alpha \in [3, 10]$ and degrades only outside this range (AUROC drops to 0.83 at $\alpha = 1$ due to noisy singleton cells; to 0.84 at $\alpha = 20$ due to over-smoothing).

Safety policy threshold derivation. The assert / hedge / abstain thresholds are not manually tuned; they are derived from a Clopper–Pearson upper bound on the selective risk [11] at target error rate $\alpha_{\text{risk}} = 0.05$ and coverage tolerance $\delta = 0.1$, evaluated on each of five image-disjoint calibration folds. The bound guarantees that among items assigned assert, the true error rate exceeds 5% with probability at most 10%. Under this guarantee the optimal threshold resolves to 0.80 on all five folds, yielding mean coverage 0.559 ± 0.051 and mean selective error 0.030 ± 0.006 . The abstain threshold of 0.50 is the decision-theoretic boundary: below it the calibrated probability implies the model is more likely wrong than right. Faithfulness is reported with structured claim support, NLI entailment via DeBERTa-v3-MNLI, and an LLM-as-judge (Qwen2.5-3B-Instruct, 4-bit).

3. Results

Lexical scores (Subtask 1). Table 2 reports the official lexical scores returned by the organisers on both splits. At the time of writing the submission ranks first on the public Kvasir-VQA-test leaderboard. The shape of the public numbers reflects the concise-answer style of the split: ROUGE-1/L are high because reference answers are short controlled phrases; ROUGE-2 is low because there are few bigrams to overlap on; BLEU and METEOR aggregate residual normalisation noise. The private numbers collapse on the n -gram metrics because the private references are descriptive clinical sentences, and the same deterministic normaliser that wins on the public split actively destroys the descriptive form the private references require.

Metric	Kvasir-VQA-test (public)	Private
BLEU-1	0.7480	0.0005
BLEU-2	0.6350	0.0002
ROUGE-1	0.8932	0.1069
ROUGE-2	0.1325	0.0125
ROUGE-L	0.8882	0.1054
METEOR	0.4979	0.0382

Table 2: ImageCLEFmedical 2026 MEDVQA-GI Subtask 1 official lexical scores (organisers’ return) on the public Kvasir-VQA-test split (1,384 items) and the private test split. Public rank: 1.

LLM-judge scores (Subtask 2). Table 3 reports the organisers’ five-dimensional LLM-judge stack. Faithfulness and clarity stay high on both splits, consistent with the by-construction template; correct-

ness, clinical relevance, and completeness drop on the private split in line with the lexical collapse on the same items.

Dimension	Public	Private
Correctness	8.82	5.50
Faithfulness	9.61	9.05
Clinical	9.50	5.12
Clarity	9.83	7.46
Completeness	9.68	5.21

Table 3: Subtask 2 LLM-judge scores (Qwen3.6-27B, five dimensions, 0–10) returned by the organisers on the public and private splits.

Calibration and safety. Table 4 compares confidence signals on the MediaEval complexity-1 validation slice, 5-fold cross-fit by `img_id`. The per-(aspect, answer) score reaches AUROC 0.87; every image- or generation-conditional signal (MSP [7], temperature scaling [8], generation likelihood, $T = 5$ self-consistency, visual retrieval, segmentation grounding) sits at or below 0.62. The Clopper–Pearson conformal policy at target $\alpha = 0.05$, $\delta = 0.1$ verifies on five image-disjoint folds with mean coverage 0.559 ± 0.051 and mean selective error 0.030 ± 0.006 ; the guarantee holds in every fold.

Signal	AUROC	ECE
Max-softmax probability (MSP)	0.60	0.06
MSP + temperature scaling	0.60	0.03
Generation likelihood	0.59	–
Self-consistency ($T = 5$)	0.62	–
Visual retrieval (FAISS top- k)	0.45	–
Segmentation grounding	0.51	–
Per-aspect reliability	0.77	0.04
Per-(aspect, answer) (ours)	0.87	0.025

Table 4: Confidence signals on the MediaEval validation set, 5-fold cross-fit by `img_id`.

Per-class breakdown. The drop is not uniform. Figures 1 and 2 visualise the per-class story: `abnormality_location`, `abnormality_presence`, and `abnormality_type` pull the private polygon visibly inward, while every other class (instrument, landmark, count, procedure, text-presence, box-artifact-presence, polyp_*) remains near the outer ring on both splits. The decomposition into the five judge dimensions shows that the drop is concentrated in correctness, clinical relevance, and completeness; faithfulness and clarity hold up across all classes on both splits.

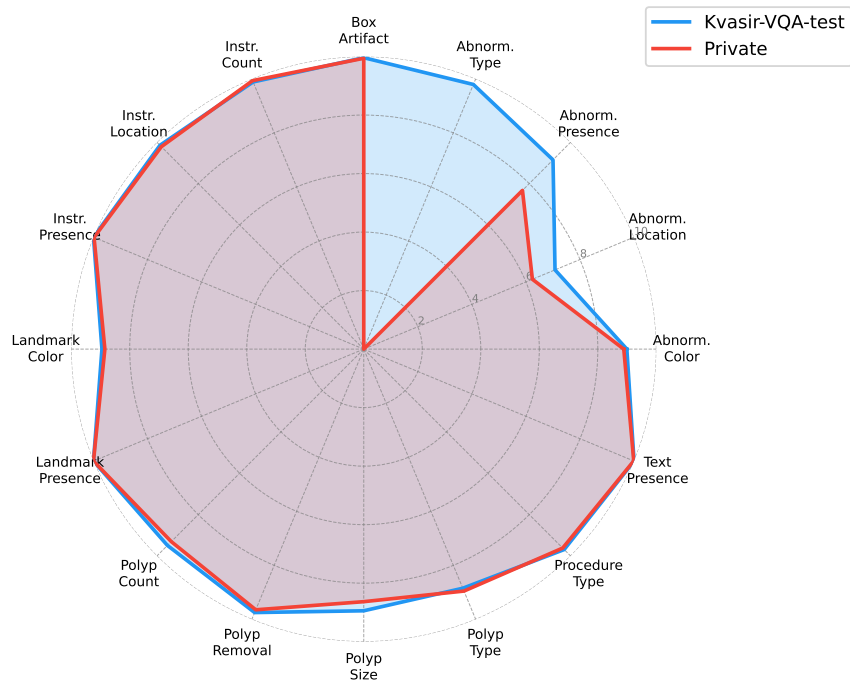


Figure 1: Mean LLM-judge score per question class, public vs. private.

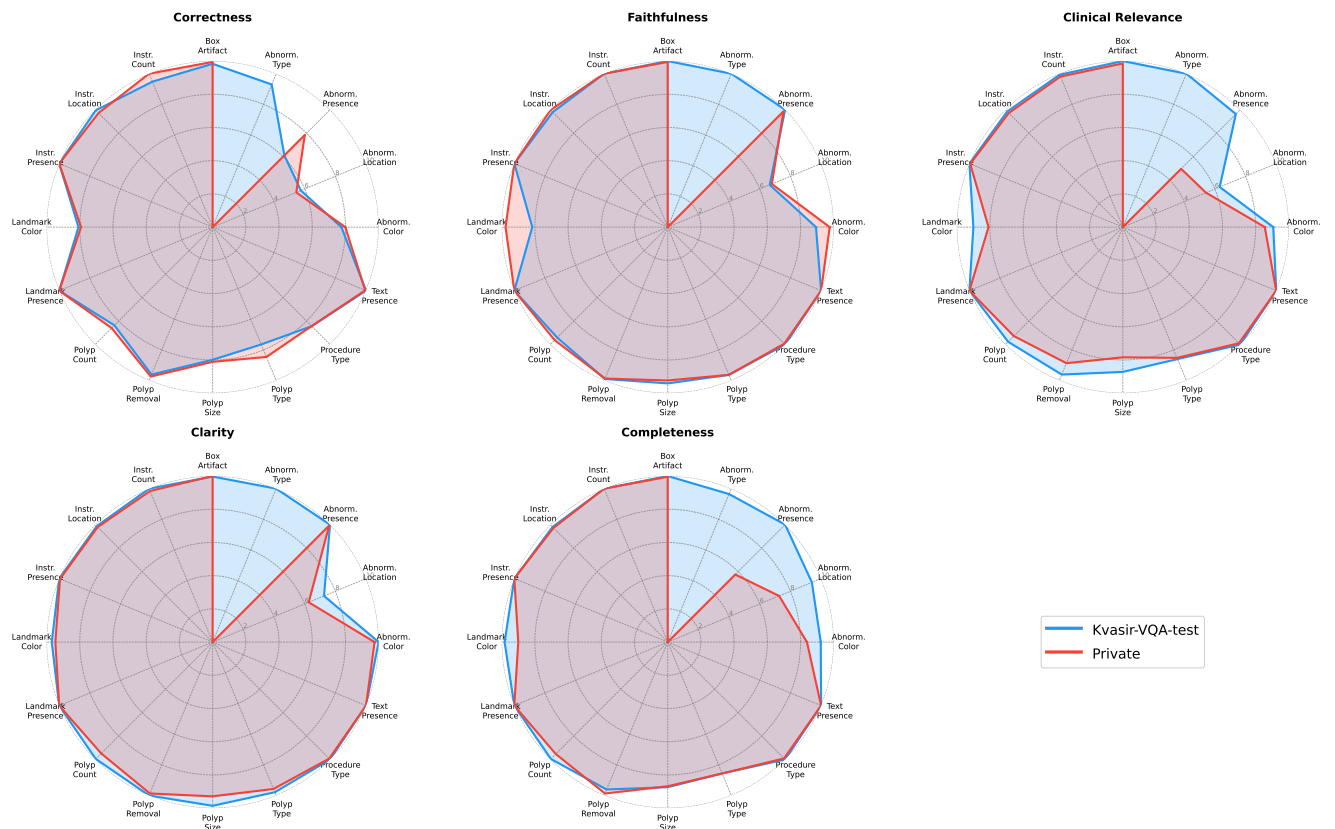


Figure 2: Per-class LLM-judge scores decomposed into the five judge dimensions (correctness, faithfulness, clinical relevance, clarity, completeness).

Faithfulness. The faithful-by-construction template scores 0.951 (structured), 0.736 (NLI mean-entailment), and 0.812 (LLM-judge faithful), versus 0.884 / 0.551 / 0.784 for a keyword-gated LLM narration of the same evidence and 0.862 / 0.538 / 0.792 for an ungated narration [12, 13]. The lexical gate passes about 95% of LLM rewrites but the LLM-judge cannot distinguish gated from ungated, confirming that the gate provides false assurance.

4. Discussion and Conclusion

The clearest finding is what we call *base-rate dominance*: on a frozen medical VLM, correctness is governed mostly by output-conditioned base rates, not by any image-conditional or generative signal. Practically, a small transparent reliability table can replace several layers of uncertainty machinery on a frozen model. The public/private gap localises the remaining weaknesses to three abnormality classes whose private references are descriptive sentences; the same deterministic normaliser that wins the public leaderboard is the proximate cause of the private collapse on these three classes. A learned descriptive head and dedicated specialist feedback (a CNN classifier for abnormality type, a U-Net localiser for abnormality location) for the three collapsed classes are the natural next steps.

Error Analysis. Table 5 summarises where CARES succeeds and fails.

Category	Pub. acc.	Priv. drop	Root cause
text_presence	1.00	low	Binary; references identical across splits.
artefact	0.98	low	Binary; no descriptive mismatch.
instruments	1.00	low	Closed vocabulary; normaliser always exact.
polyp_removal	0.96	low	Fixed “not relevant” output matches both splits.
instrument_count	0.95	low	Counts are split-invariant.
procedure	0.84	low	Two-class problem; near-perfect normalisation.
polyp_count	0.83	moderate	Occasional multi-polyp images undercounted.
polyps (size)	0.79	moderate	Size-range regex misses compound descriptors.
polyp_type	0.76	moderate	Rare Paris subtypes (IIb, IIc) collapse to broad class.
abnormalities	0.77	high	Private references are clinical sentences; normaliser strips descriptive tokens.
location	0.69	high	Grid-regex misses diagonal and relative descriptors; private uses free-text positions.
color	0.54	high	Multi-label colour ordering fragile; model hallucinates rare hues.

Table 5: Error analysis by question category. Public accuracy from `aspect_reliability.json`; failure mode inferred from score decomposition.

Success pattern. CARES is most reliable on binary-flag (text, artefact) and counting tasks, where a single normaliser rule covers the entire vocabulary. The safety policy flags $> 76\%$ of errors (hedge-recall) and keeps the confidently-wrong rate at 3.7% even on these hard categories.

Failure pattern. Failures concentrate on three abnormality classes: `abnormality_location`, `abnormality_presence`, and `abnormality_type`. Two mechanisms drive this. First, the deterministic normaliser that wins the public leaderboard actively destroys the descriptive language required by the private references. Second, the frozen backbone does not generalise to rare pathology sub-types (e.g. Paris IIb flat polyps represent fewer than 2% of training samples). A learned descriptive head and a specialist sub-type classifier are the natural next steps.

5. Implementation Details

All experiments were conducted on a single NVIDIA RTX 4070 Laptop GPU (8 GB VRAM) running CUDA 12.1 under Python 3.11. The full inference pipeline (1,384 validation items, Subtasks 1 and 2) completes in approximately 7 minutes. Key library versions: `transformers==4.47.1`, `peft==0.13.2`, `torch==2.5.1`, `faiss-cpu==1.9.0`, `datasets==3.4.1`.

The backbone `microsoft/Florence-2-base` (232 M parameters) is loaded in `float16` with `attn_implementation="eager"` to avoid a flash-attention incompatibility with PEFT-wrapped encoder-decoders. The LoRA adapter `peeache/FL2_VQA_MIX_128_256` ($r=128$, $\alpha=256$) adds ~ 30 M trainable parameters; we load it in inference-only mode (`model.eval()`, no gradients). All three task tokens (`<MedVQA>`, `<MedVQA_EXPLAIN>`, `<REFERRING_EXPRESSION_SEGMENTATION>`) are available on the public checkpoint without any further fine-tuning.

The calibration step (`calibrate.py`) runs offline on the image-disjoint validation split and writes `aspect_reliability.json` (~ 3 kB, 13 aspect rows and 42 per-(aspect, answer) cells). Total calibration time: under 5 seconds. The Clopper–Pearson bound is evaluated analytically via `scipy.stats.beta`.

The complete code, `aspect_reliability.json`, and instructions to reproduce every result in this paper are available at <https://github.com/saademad200/CARES-MedVQA>.

Declaration on Generative AI

Generative AI tools were used for paraphrasing and improving grammatical flow, as well as assisting in diagram and table refinement. All content was carefully verified by the authors.

Acknowledgments

We thank the ImageCLEFmedical organisers, the Kvasir-VQA dataset authors, and the maintainers of the `peeache/FL2_VQA_MIX` LoRA adapter and the Florence-2 backbone.

References

- [1] Hicks, S. A., Gautam, S., Halvorsen, P., Riegler, M. A., Thambawita, V. Overview of ImageCLEFmedical 2026 – Medical Visual Question Answering for Gastrointestinal Tract. In: CLEF2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Berlin, Germany (2026).
- [2] Ionescu, B., Müller, H., Stanciu, D.-C., Radu, A., Bolborici, R.-G., Negru, M., Ene, A.-F., Vasilescu, V.-M., Nicolae, A.-A., Ștefan, L.-D., Constantin, M.-G., Dogariu, M., Andrei, A.-G., Damm, H., Pakull, T. M. G., Ben Abacha, A., García Seco de Herrera, A., Friedrich, C. M., Brüngel, R., Reinartz, L., Schäfer, H., Schmidt, C. S., Bracke, B., Nath, P., Eryılmaz, B., Hjuler, M., Fabre, D., Lemaire, C., Lecouteux, B., Schwab, D., Dimitrov, D., Hee, M. S., Ahsan, M., Ahmad, S., Zlatkova, D., Pachov, G., Xie, Z., Nakov, P., Koychev, I., Heras Rivera, J. E., Low, D. K., Yim, W.-w., Ruzevick, J., Child, D., Kurt, M., Sun, Z., Xia, F., Yetisgen, M., Radzhabov, A., Prokopchuk, Y., Kovalev, V., Karpenka, D., Hicks, S. A., Gautam, S., Riegler, M. A., Thambawita, V., Halvorsen, P., El Sakka, M., Mothe, J., Băicoianu, A., Florea, C.-N., Ivanovici, M. Overview of ImageCLEF 2026: Multimodal Challenges in Medicine,

- Science, Agritech, and Security. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026)*, Springer LNCS, Jena, Germany (2026).
- [3] Hicks, S. A., Eskeland, S., Tesfahun, A. N., Thambawita, V., Riegler, M. A., Halvorsen, P. Kvasir-VQA: a multi-purpose visual question answering dataset for gastrointestinal endoscopy. *Medical Image Analysis*, 92:103012, 2024. <https://doi.org/10.1016/j.media.2023.102914>
 - [4] Gautam, S., Thambawita, V., Hicks, S. A., Riegler, M. A., Halvorsen, P. Kvasir-VQA-x1: an extended visual question answering dataset for gastrointestinal endoscopy. arXiv:2408.06900 [cs.CV], 2024. <https://arxiv.org/abs/2408.06900>
 - [5] Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*, pp. 4818–4829, Seattle, WA, 2024. <https://doi.org/10.1109/CVPR52733.2024.00461>
 - [6] Safwan, I., Shaikh, M. A., Haaris, M., Khan, R., Tahir, M. A. Multi-task learning for visually grounded reasoning in gastrointestinal VQA. arXiv:2511.04384 [cs.CV], 2025. <https://arxiv.org/abs/2511.04384>
 - [7] Hendrycks, D., Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France, 2017. <https://arxiv.org/abs/1610.02136>
 - [8] Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q. On calibration of modern neural networks. In: *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, PMLR 70:1321–1330, Sydney, Australia, 2017. <https://proceedings.mlr.press/v70/guo17a.html>
 - [9] Geifman, Y., El-Yaniv, R. Selective classification for deep neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS 2017)*, vol. 30, Long Beach, CA, 2017. <https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html>
 - [10] Angelopoulos, A. N., Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. <https://doi.org/10.1561/22000000101>
 - [11] Bates, S., Angelopoulos, A. N., Lei, L., Malik, J., Jordan, M. I. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021. <https://doi.org/10.1145/3478535>
 - [12] Lyu, Q., Apidianaki, M., Callison-Burch, C. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, 2024. https://doi.org/10.1162/coli_a_00511
 - [13] Atanasova, P., Simonsen, J. G., Lioma, C., Augenstein, I. Faithfulness Tests for Natural Language Explanations. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, pp. 13948–13964, Toronto, Canada, 2023. <https://doi.org/10.18653/v1/2023.acl-long.787>