Medico 2026: Visual Question Answering for Gastrointestinal Imaging

Sushant Gautam[1,3,*], Vajira Thambawita[1], Michael Riegler[2], Pål Halvorsen[1,3] and Steven Hicks[1]

[1]SimulaMet - Simula Metropolitan Center for Digital Engineering, Oslo, Norway
[2]Simula Research Laboratory, Oslo, Norway
[3]OsloMet - Oslo Metropolitan University, Oslo, Norway

Abstract

The Medico 2026 challenge addresses Visual Question Answering (VQA) for Gastrointestinal (GI) imaging, organized as part of the MediaEval task series. The task focuses on developing Explainable Artificial Intelligence (XAI) models that can answer clinically relevant questions based on GI endoscopy images while providing coherent and clinically grounded explanations. Building on the success of previous editions and closely aligned with the ImageCLEFmedical-MEDVQA 2026 initiative, the challenge introduces two subtasks: (1) answering diverse types of visual questions using the Kvasir-VQA-x1 dataset, and (2) generating explainable and safe multimodal reasoning to support clinical decision-making. The Kvasir-VQA-x1 dataset, containing more than 150,000 clinically relevant question–answer (QA) pairs, serves as the benchmark for evaluation. In addition to quantitative metrics and expert-reviewed explainability assessments, the 2026 edition introduces evaluation criteria targeting behavioral safety, discouraging undesirable model behaviors such as overconfident answers, misleading justifications, or clinically inappropriate reasoning. The task aims to advance trustworthy and interpretable Artificial Intelligence (AI) decision support for GI diagnostics. Instructions, data access, and participation guidelines are available in the official repository: github.com/simula/MediaEval-Medico-2026

1. Introduction

Gastrointestinal (GI) diseases represent a major global health burden, where accurate interpretation of endoscopy findings is critical for diagnosis and treatment planning [1, 2]. While AI-driven decision support systems [3, 4] have demonstrated strong performance in GI image analysis, their clinical adoption remains limited by insufficient explainability, safety concerns, and a lack of alignment with clinical reasoning [5, 6]. Building on the success of previous Medico challenges at MediaEval and closely aligned with the ImageCLEFmedical-MEDVQA 2026 initiative, the 2026 edition[1] introduces *Medico 2026: Visual Question Answering for Gastrointestinal Imaging*.

Medical VQA is a rapidly growing research area that combines computer vision and natural language processing to answer clinically meaningful questions derived from medical images [5]. However, existing approaches often prioritize answer accuracy without sufficiently addressing explanation quality, safety, or clinical consistency [5, 6]. To address this, the Medico 2026 challenge emphasizes not only correct answers but also multimodal explanations that combine textual and visual evidence and adhere to medical best practices. In addition, the task introduces evaluation criteria targeting behavioral safety, discouraging undesirable model behaviors such as overconfident answers, misleading justifications, or clinically inappropriate reasoning. The challenge provides a benchmark dataset of GI images and associated VQA annotations, enabling

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
sushant@simula.no (S. Gautam); vajira@simula.no (V. Thambawita); michael@simula.no (M. Riegler); paalh@simula.no (P. Halvorsen); steven@simula.no (S. Hicks)

[1]https://multimediaeval.github.io/editions/2026/tasks/medico

rigorous evaluation of AI models. By integrating multimodal data, explainability metrics, and safety evaluation, we aim to advance research in trustworthy AI and increase the potential for clinical adoption.

We define two main subtasks for this year's challenge. Subtask 2 builds on Subtask 1, meaning Subtask 1 must be completed in order to participate in Subtask 2.

- **Subtask 1:** Medical Image Question Answering in GI Endoscopy
  This subtask focuses on developing models that accurately answer clinically relevant questions based on GI endoscopy images using the Kvasir-VQA-x1 dataset, which contains more than 150,000 QA pairs. The dataset is derived from established GI endoscopy collections [7] and covers a wide range of anatomical regions, pathological findings, and medical instruments. Questions span multiple categories, including Yes/No, Single-Choice, Multiple-Choice, Color-Related, Location-Related, and Numerical Count, requiring joint reasoning over visual and textual information. Performance will be assessed using Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-1/2/L, and Metric for Evaluation of Translation with Explicit ORdering (METEOR).
- **Subtask 2:** Explainable and Safe Multimodal Reasoning for GI VQA
  This subtask extends Subtask 1 by requiring models to provide coherent multimodal explanations that justify their answers. Explanations must combine textual reasoning with visual evidence—such as highlighted image regions—in a manner aligned with clinical reasoning [8, 9, 10, 11, 12]. In addition to interpretability, this subtask introduces a dedicated safety layer that evaluates model behavior across clinical contexts. Models are assessed for undesirable behaviors, including overconfidence, misleading explanations, or non-compliance with established medical best practices. To support retrieval-augmented reasoning, participants may leverage a curated database of verified endoscopy resources provided as part of the challenge. Submissions will be evaluated across multiple dimensions, including answer correctness, explanation clarity, coherence, medical relevance, consistency with visual evidence, and behavioral safety. The evaluation will emphasize the overall quality, interpretability, medical grounding, and safety of the explanations, assessing how effectively they support clinical decision-making.

For AI systems to be integrated into clinical workflows, they must be transparent, interpretable, and safe. In GI imaging, deep learning models have achieved promising results for classification and detection tasks, yet their black-box nature limits trust among clinicians. Medical professionals require explanations that clearly connect visual evidence to clinical conclusions.

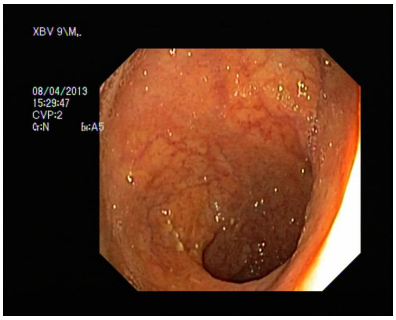
Medical VQA offers a natural interface for explainable decision support, enabling clinicians to ask structured questions and receive interpretable responses. Nevertheless, many existing VQA models provide answers without sufficient justification or safeguards against unsafe reasoning. Medico 2026 addresses these limitations by explicitly integrating explainability and safety into both task design and evaluation. By encouraging multimodal explanations and clinically consistent behavior, the challenge aims to advance AI systems that support, rather than replace, clinical expertise.

2. Data

The Medico 2026 challenge builds on the Kvasir-VQA-x1 dataset [13], an extensive extension of Kvasir-VQA [7]. It contains 6,500 gastrointestinal (GI) endoscopic images from HyperKvasir [14]

and Kvasir-Instrument [15], paired with more than 150,000 question–answer (QA) pairs stratified by reasoning complexity. The dataset is curated with clinical input to ensure medical relevance and correctness.

Table 1
An example image with one representative question–answer pair from each complexity level in the Kvasir-VQA-x1 dataset. Each image in the dataset may have multiple QA pairs at each level.



Complexity	Question	Answer	Question Class
1	Which anatomical landmark is visible in the image?	No identifiable anatomical landmark present	landmark_location
2	What procedure is depicted in the image and what colors are associated with the abnormality?	Evidence of colonoscopy findings with pink and red mucosal lesions	procedure_type, abnormality_color
3	Are there any anatomical landmarks visible, what type of polyps are present, and what colors are the observed abnormalities?	No anatomical landmarks identified, no polyps observed, and multiple abnormalities with pink and red coloration.	landmark_presence, polyp_type, abnormality_color

Each image is linked to multiple QA entries generated by merging one to three atomic pairs using the Qwen3-30B-A3B model [16]. The resulting questions are fluent and annotated with a `complexity` score (1–3) and a `question_class` label describing the clinical focus, such as `polyp_type`, `instrument_presence`, or `finding_count`.

Question complexity is categorized into three levels: single, dual, and triple atomic reasoning—representing approximately 34.4%, 32.8%, and 32.8% of the data, respectively. Each QA pair is further annotated with one or more `question_class` labels for fine-grained analysis across clinical aspects including pathology, anatomy, procedure, and visual findings.

Kvasir-VQA-x1 is publicly available at https://huggingface.co/datasets/SimulaMet/Kvasir-VQA-x1. Each entry includes `img_id`, `complexity`, `question`, `answer`, `original` (atomic components), and `question_class`, and the dataset is split into training and test sets for reproducibility. Only original images are provided; however, weak augmentations (e.g., rotation, color jitter, cropping) are encouraged during fine-tuning.

3. Evaluation

The Medico 2026 challenge evaluates both the **accuracy** and **clinical interpretability** of medical VQA models, emphasizing not only correct answers but also their **relevance, explanatory quality, and safety** in GI diagnostics [17].

3.1. Subtask 1: Medical Image Question Answering in GI Endoscopy

This subtask evaluates how effectively models answer clinically relevant GI questions from medical images, focusing on both predictive accuracy and language accuracy. Performance is measured using BLEU, ROUGE-1/2/L, and METEOR, assessing alignment with reference responses. A small portion of the Kvasir-VQA-x1 test set will be used for public score calculation in Subtask 1. The resulting scores will be automatically displayed on the leaderboard upon model submission, allowing participants to monitor their model's performance throughout the competition.

Detailed model assessment occurs at three levels: (1) *overall performance*, aggregating all question types; (2) *category-level analysis*, across question types (e.g., polyp type, instrument presence) visualized via radar plots; and (3) *complexity-level evaluation*, distinguishing factual

(Level 1), inferential (Level 2), and higher-order reasoning (Level 3) questions. This structured framework enables a comprehensive evaluation of both correctness and clinical reasoning—key considerations for reliable deployment in medical settings.

3.2. Subtask 2: Explainable and Safe Multimodal Reasoning

Subtask 2 extends Subtask 1 by requiring participants to provide coherent multimodal explanations that justify their answers. While Subtask 1 focuses on automated metrics for answer correctness, Subtask 2 emphasizes the quality, interpretability, medical grounding, and safety of the model's explanations.

Each submission must combine textual reasoning with visual evidence—such as highlighted image regions—that connects the predicted answer to diagnostic evidence in a manner aligned with clinical reasoning. To support retrieval-augmented reasoning, participants may leverage a curated database of verified endoscopy resources provided as part of the challenge.

In addition to interpretability, this subtask introduces a dedicated safety layer that evaluates model behavior across clinical contexts. Models are assessed for undesirable behaviors, including overconfidence, misleading explanations, or non-compliance with established medical best practices. Safety-oriented criteria evaluate whether model outputs demonstrate appropriate uncertainty, factual correctness, and adherence to clinical best practices.

Submissions will be evaluated across multiple dimensions, including answer correctness, explanation clarity, coherence, medical relevance, consistency with visual evidence, and behavioral safety. The detailed evaluation protocol will be released after the competition concludes, depending on the diversity of submitted systems. Participants are encouraged to produce explanations that are coherent, medically accurate, safe, and well aligned with visual evidence to achieve strong performance.

For the final phase of the challenge, a private challenge set, unseen by participants, will be used to assess model performance for both Subtask 1 and Subtask 2. This private set will contain completely new images not included in the Kvasir-VQA-x1 dataset, ensuring that models are evaluated on truly unseen data. The evaluation on this private set will determine the final rankings and will be conducted by the organizers. The evaluation results will be communicated to participants after the submission period concludes to ensure a fair and unbiased comparison of systems.

4. Discussion and Outlook

The Medico 2026 challenge marks an important step toward bridging the gap between powerful deep learning models and their practical adoption in clinical settings. By focusing on explainable and safe VQA for GI imaging, this task promotes the development of interpretable AI models that not only generate accurate responses but also provide transparent justifications aligned with medical reasoning while adhering to safety best practices.

Participants are encouraged to innovate beyond traditional accuracy metrics and embrace multimodal explainability and behavioral safety as core components of their solutions. The availability of the Kvasir-VQA-x1 dataset, tailored for this task, will support reproducible research and enable robust benchmarking. Looking ahead, we anticipate that methods developed for Medico 2026 will inspire broader applications of explainable and safe AI in other medical domains. By fostering interdisciplinary collaboration between the AI and medical communities, this challenge aims to pave the way for clinically viable AI tools that are both trusted and actionable in real-world healthcare scenarios.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude in order to perform *grammar and spelling checks, paraphrasing and rewording, and improving the writing style.* After using this tool/service, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Singh, Global burden of five major types of gastrointestinal cancer, Gastroenterology Review/Przegląd Gastroenterologiczny 19 (2024).

[2] R. Wang, Z. Li, S. Liu, D. Zhang, Global, regional, and national burden of 10 digestive diseases in 204 countries and territories from 1990 to 2019, Frontiers in public health 11 (2023) 1061453.

[3] H. Ali, M. A. Muzammil, D. S. Dahiya, F. Ali, S. Yasin, W. Hanif, M. K. Gangwani, M. Aziz, M. Khalaf, D. Basuli, et al., Artificial intelligence in gastrointestinal endoscopy: a comprehensive review, Annals of gastroenterology 37 (2024) 133.

[4] M. A. Berbís, J. Aneiros-Fernández, F. J. M. Olivares, E. Nava, A. Luna, Role of artificial intelligence in multidisciplinary imaging diagnosis of gastrointestinal diseases, World journal of gastroenterology 27 (2021) 4395.

[5] K. Borys, Y. A. Schmitt, M. Nauta, C. Seifert, N. Krämer, C. M. Friedrich, F. Nensa, Explainable AI in medical imaging: An overview for clinical practitioners – Beyond saliency-based XAI approaches, Eur. J. Radiol. 162 (2023). doi:10.1016/j.ejrad.2023.110786.

[6] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, P. Lambin, Transparency of deep neural networks for medical image analysis: A review of interpretability methods, Comput. Biol. Med. 140 (2022) 105111. doi:10.1016/j.compbiomed.2021.105111.

[7] S. Gautam, A. M. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, Kvasir-vqa: A text-image pair gi tract dataset, in: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications, 2024, pp. 3–12.

[8] D. Muhammad, M. Bendechache, Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis, Comput. Struct. Biotechnol. J. 24 (2024) 542–560. doi:10.1016/j.csbj.2024.08.005.

[9] X. Gai, C. Zhou, J. Liu, Y. F. (xn-27q. xn-6xw.), J. Wu, Z. Liu, MedThink: A Rationale-Guided Framework for Explaining Medical Visual Question Answering, ACL Anthology (2025) 7438–7450. doi:10.18653/v1/2025.findings-naacl.415.

[10] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal Explanations: Justifying Decisions and Pointing to the Evidence, IEEE Computer Society, 2018. doi:10.1109/CVPR.2018.00915.

[11] A. M. Storås, M. Dreyer, F. Pahde, S. Lapuschkin, W. Samek, P. Halvorsen, T. de Lange, Y. Mori, A. Hann, T. M. Berzin, S. Parasa, M. A. Riegler, Exploring the clinical value of concept-based AI explanations in gastrointestinal disease detection, Sci. Rep. 15 (2025) 1–11. doi:10.1038/s41598-025-14408-y.

[12] F. Dahan, J. H. Shah, R. Saleem, M. Hasnain, M. Afzal, T. M. Alfakih, A hybrid XAI-driven deep learning framework for robust GI tract disease diagnosis, Sci. Rep. 15 (2025) 1–18. doi:10.1038/s41598-025-07690-3.

[13] S. Gautam, M. A. Riegler, P. Halvorsen, Kvasir-VQA-x1: A Multimodal Dataset for Medical Reasoning and Robust MedVQA in Gastrointestinal Endoscopy, arXiv (2025). doi:10.48550/arXiv.2506.09958. arXiv:2506.09958.

[14] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, et al., Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Scientific data 7 (2020) 283.

[15] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. De Lange, P. T. Schmidt, H. D. Johansen, et al., Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: MultiMedia Modeling: 27th International Conference,

MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27, Springer, 2021, pp. 218–229.

[16] A. Yang, A. Li, B. Yang, et al., Qwen3 Technical Report, arXiv (2025). doi:10.48550/arXiv.2505.09388. arXiv:2505.09388.

[17] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, S. Parasa, On evaluation metrics for medical applications of artificial intelligence, Scientific reports 12 (2022) 5979.