Overview of ImageCLEF 2026: Multimodal Challenges in Medicine, Science, Agritech, and Security

Bogdan Ionescu[1], Henning Müller[2], Dan-Cristian Stanciu[1], Andrei Radu[1], Radu-George Bolborici[1], Marian Negru[1], Alexandru-Florin Ene[1], Vlad-Mihai Vasilescu[1], Ana-Antonia Nicolae[1], Liviu-Daniel Ştefan[1], Mihai-Gabriel Constantin[1], Mihai Dogariu[1], Alexandra-Georgiana Andrei[1], Hendrik Damm[3], Tabea M. G. Pakull[4,3], Asma Ben Abacha[5], Alba García Seco de Herrera[6], Christoph M. Friedrich[3], Raphael Brüngel[3], Lea Reinartz[3], Henning Schäfer[4,3], Cynthia Sabrina Schmidt[4], Benjamin Bracke[3], Praveen Nath[3], Bahadır Eryılmaz[7], Maja Hjuler[8], Diandra Fabre[8], Claire Lemaire[8], Benjamin Lecouteux[8], Didier Schwab[8], Dimitar Dimitrov[9], Ming Shan Hee[10], Momina Ahsan[10], Sarfraz Ahmad[10], Dimitrina Zlatkova[9], Georgi Pachov[9], Zhuohan Xie[10], Preslav Nakov[10], Ivan Koychev[9], Juampablo E. Heras Rivera[11], Daniel K. Low[11], Wen-wai Yim[5], Jacob Ruzevick[11], Daniel D. Child[11], Mehmet Kurt[11], Zhaoyi Sun[11], Fei Xia[11], Meliha Yetisgen[11], Ahmedkhan Radzhabov[12], Yuri Prokopchuk[12], Vassili Kovalev[12], Dzmitry Karpenka[12], Steven A. Hicks[13], Sushant Gautam[13], Michael A. Riegler[13], Vajira Thambawita[13], Pål Halvorsen[13], Mohammad El Sakka[14], Josiane Mothe[14], Alexandra Băicoianu[15], Corneliu-Nicolae Florea[1,15], and Mihai Ivanovici[15]

[1] National University of Science and Technology Politehnica Bucharest, Romania
bogdan.ionescu@upb.ro
[2] University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[3] University of Applied Sciences and Arts Dortmund, Germany
[4] Institute for Transfusion Medicine, University Hospital Essen, Germany
[5] Microsoft, USA
[6] University of Distance Education (UNED), Spain
[7] Institute for Artificial Intelligence in Medicine, University Hospital Essen
[8] Université Grenoble Alpes, LIG, France
[9] Sofia University "St. Kliment Ohridski", Bulgaria
[10] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[11] University of Washington, USA
[12] National Academy of Science of Belarus, Minsk, Belarus
[13] SimulaMet, Norway
[14] Université de Toulouse, France
[15] Transilvania University of Brasov, Romania

Abstract. This paper presents an overview of the ImageCLEF 2026 lab, organized within the Conference and Labs of the Evaluation Forum – CLEF Labs 2026 [17]. This year's 24th edition of the evaluation event consists of five main tasks: (i) the *Medical* task, comprised of four sub-tasks, consisting of various challenges in the medical field, like *Concept Detection and Caption Prediction* on real and synthetic medical images, evaluation of *Veracity and Privacy of GAN-generated 3D*

CT scans, *Visual Question Answering* and explainability in gastrointestinal endoscopy images, and a new *MEDIQA-CORE* task on Brain Tumor Subtype Classification and Report Discrepancy Summarization on real-life medical data, (ii) the *ToPicto* task, focusing on aiding communication using text-to-pictogram translation and pictogram sequence prediction, (iii) a *Multimodal Reasoning* task on multi-lingual, multi-discipline question answering on multiple modalities, and two new tasks on (iv) *Deepfake Detection and Generation* in images and audio and (v) the *AI4Agriculture* task, focused on Viticultultire Potential prediction and Crop Identification. This edition of ImageCLEF attracted a total of 307 teams registered for all the different tasks. This resulted in 1,284 total runs across all competitions, and a total of 74 working notes papers submitted, making this year's one of the most successful editions.

Keywords: Medical Image Processing · Image Caption Analysis · Medical Concept Prediction · Question Answering · Generative Adversarial Networks · Synthetic Data Generation · Image Segmentation · Pictogram Communication · Multilingual · Image Retrieval · Deepfake · Agritech

1 Introduction

Since its establishment in 2003 [8], ImageCLEF[16] has become one of the leading evaluation campaigns in the field, fostering research on the indexing and retrieval of visual content as well as enabling access to large-scale data collections across diverse domains. This paper provides an overview of the 2026 edition of the ImageCLEF lab, organized as part of the Conference and Labs of the Evaluation Forum (CLEF) 2026 [16]. This edition of ImageCLEF consists of five tasks: ImageCLEFMedical (with 5 sub-tasks: Caption, GANs, MEDIQA-CORE Task 1 and Task 2, and MEDVQA), ToPicto, MultimodalReasoning, Deepfake and AI4Agriculture. The automatic evaluation for most of the tasks was carried out on our Ai4Media benchmarking platform[17].

2 Overview of Tasks and Participation

ImageCLEF 2026 consists of five tasks, covering a diverse range of applications:

- **ImageCLEFmedical**. Since 2004, the ImageCLEF benchmarking initiative has incorporated medical tasks. However, due to having little interaction between tasks, the medical tasks were merged into a single overarching medical task composed of multiple subtasks, each addressing a distinct problem, from 2019 [21] onward. This restructuring enabled greater synergy across the different domains. The medical task consists of five subtasks, described below:
  - *Caption*: The ImageCLEFmedical Caption task [9] continued the long-running benchmark on radiology image understanding by combining UMLS concept detection and medical caption generation. The 2026 edition introduced synthetic-data extensions for both concept detection and

[16] `http://www.imageclef.org/`
[17] `https://ai4media-bench.aimultimedialab.ro/`

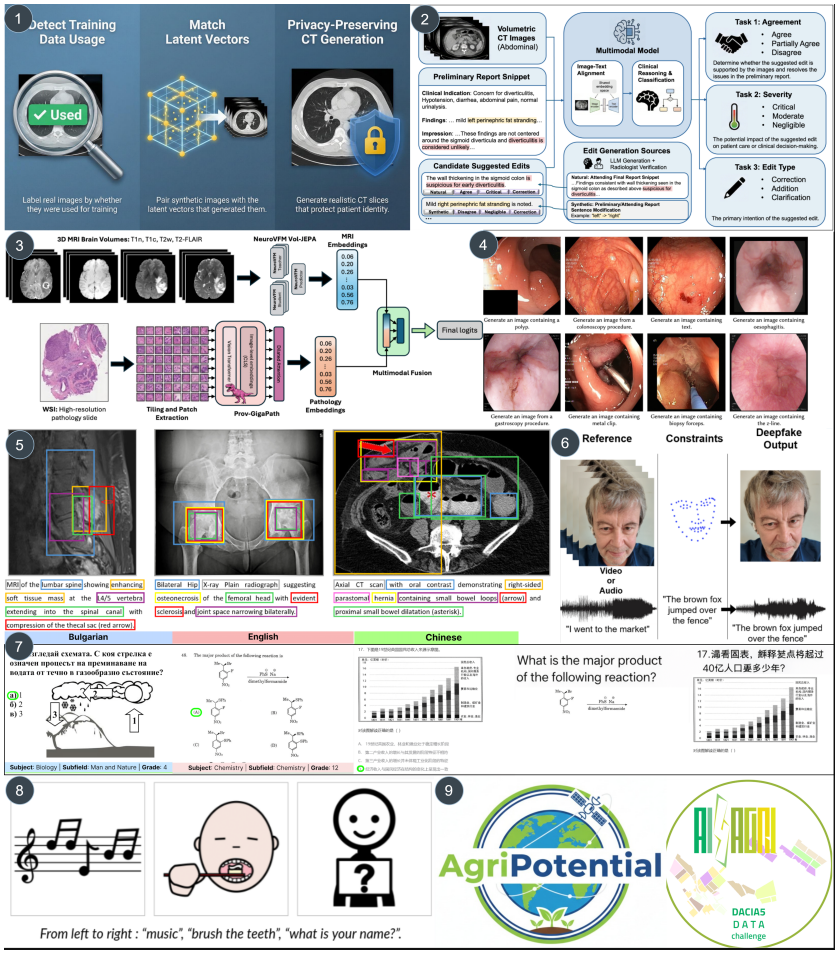


Fig. 1: Representative images for all ImageCLEF Tasks: (1) GANs, (2) MEDIQA-CORE Task 2, (3) MEDIQA-CORE Task 1, (4) MEDVQA-GI, (5) Caption, (6) Deepfake, (7) Multimodal Reasoning, (8) ToPicto, (9) Ai4Agriculture.

caption prediction, enabling the evaluation of systems under lexical and stylistic distribution shifts, and added a manually evaluated explainability subtask in which participants provided human-interpretable explanations for generated captions.

- *MEDIQA-CORE Task 1* [5], organized within the MEDIQA shared tasks[18] and ImageCLEF 2026, focused on multimodal brain tumor subtype classification under incomplete-modality settings. The benchmark combined multi-sequence brain MRI, histopathology whole-slide images, and radiology reports, while also including cases with missing histopathology data to evaluate robustness to unavailable modalities. Participants ad-

[18] `https://sites.google.com/view/mediqa-shared-tasks`

dressed three classification tasks: molecular subtype grouping, WHO grade prediction, and low-grade versus high-grade glioma classification.
    - *MEDIQA-CORE Task 2* [6], focuses on multimodal radiology report review and discrepancy assessment. Using expert-annotated abdominal CT studies that mirror trainee-to-attending report revision workflows, each case included a volumetric abdominal CT examination, a preliminary radiology report, and a candidate edit. Participants developed systems to determine whether edits were supported by imaging findings, assess the clinical severity of discrepancies, and classify the edit type.
    - *GANs*: The fourth edition of the ImageCLEFmedical GANs task [1] builds on the findings learned from the previous editions [3, 4, 2] and continues to investigate privacy and security concerns in GAN-generated synthetic medical images. This edition introduces three complementary subtasks: detecting which real images were used to train a GAN (Subtask 1), identifying the latent vector that produced each synthetic image (Subtask 2), and generating privacy-preserving CT slices evaluated through four independent privacy attacks (Subtask 3).
    - *MEDVQA*: The MEDVQA challenge [15] answers the need for research in Analysis of gastrointestinal (GI) images by introducing a multimodal task that combines image and text data for visual question answering (VQA), targeting the field of GI endoscopy. The challenge includes two subtasks: (1) answering clinical questions associated with specific images and (2) generating synthetic GI images based on prompts describing anatomical landmarks, visual features, and other relevant findings.
- The **Deepfake** task, in its first edition, aims to bridge the gap between experiments and real-life models in regards to the generalization of deepfake detectors. The task is split into Deepfake Generation and Deepfake Detection, for Images and Audio. The task's evaluation procedure has a cyclical nature: the data in the Generation subtask is given as an input to the participants in the Detection task, evaluating deepfake detection models as well as detection evasion and quality on the generated data.
- **ToPicto**. The ImageCLEFtoPicto task, in its third edition, continues to advance research on the automatic conversion of text into meaningful sequences of pictograms designed to support Alternative and Augmentative Communication (AAC). The 2026 edition introduces two significant changes relative to the previous editions. First, the source language shifts from French to English, broadening the task's reach and posing new linguistic, syntactic, and semantic challenges. Second, a new next-pictogram prediction task is introduced alongside the established translation task.
- **MultimodalReasoning** The second edition of the Multimodal Reasoning task benchmarks open-weight and open-source vision-language systems on multilingual academic examination questions. The task covers both multiple-choice question answering and open-ended answer generation from question images, including examples with and without visual elements such as diagrams, graphs, tables, and figures. This year's edition extends the previous benchmark with new open-ended questions and repurposed multiple-choice

questions, allowing systems to be evaluated both on answer selection and on direct answer generation without visible answer options.
– **Ai4Agri**. The Ai4Agri task is composed of two subtasks related to the innovative use of Artificial Intelligence and Remote Sensing in the field of Smart Agriculture: *AgriPotential: Viticulture Edition*, consisting of predicting the agricultural potential of land for grapevine cultivation using high-resolution Sentinel-2 image time series and *DACIA 5*, on agricultural crops identification in remotely sensed multispectral images.

Table 1: Key figures regarding participation in ImageCLEF 2026.

Task	Groups that submitted results	Submitted runs	Submitted working notes
Deepfake	26	205	16
Caption	20	135	16
MultimodalReasoning	13	147	11
ToPicto	13	94	9
GANs	10	51	6
MEDIQA-CORE Task 1	5	95	4
MEDIQA-CORE Task 2	4	92	4
MEDVQA	9	112	4
Ai4Agri	6	184	4

3 The Caption Task

The ImageCLEFmedical Caption task [9] addresses automatic understanding of radiology images through concept detection, caption generation, and explainability. The 2026 edition extends the established benchmark with two synthetic-data subtasks in addition to the two core real-data subtasks. This design makes it possible to evaluate not only performance on real radiology captions, but also robustness to distribution shifts introduced by synthetic captions. A fifth subtask focuses on explainability and asks participants to provide human-interpretable justifications for caption outputs on a small subset of test images.

3.1 Task Setup

The challenge was organized into five subtasks. The first two subtasks used real radiology images: *Main Concept Detection*, where systems predicted UMLS Concept Unique Identifiers (CUIs), and *Main Caption Prediction*, where systems generated one caption per image. The third and fourth subtasks repeated the same prediction problems on the synthetic-caption dataset: *Synthetic Concept Detection* and *Synthetic Caption Prediction*. The final subtask, *Explainability*, required participants to provide explanations for caption outputs and was evaluated manually by a radiologist.

For the concept detection subtasks, submissions consisted of image identifiers paired with semicolon-separated CUIs and were evaluated primarily using F1-score. Caption prediction submissions consisted of one generated caption per image and were evaluated with a composite score combining relevance and factuality metrics. Relevance was assessed using text-based and image-text similarity metrics, while factuality was assessed using UMLS concept overlap and factual consistency measures. The explainability task used manual Likert-scale assess-

ment over caption quality, visualisation quality, methodology, and overall clinical preference.

3.2 Dataset

The 2026 release contains two related benchmark datasets. The main dataset is based on radiology images and captions from PubMed Central Open Access articles [27] and was prepared through the yearly ImageCLEFmedical data pipeline. It contains 131 853 images in total, split into 97 364 training images, 19 240 validation images, and 15 249 test images, with 2 646 unique CUIs. The data preparation process included split handover from the previous edition, regenerated yearly identifiers, missing-image pruning, zero-concept removal, and release of CUI-to-name mappings.

The synthetic-caption dataset contains 131 723 images, split into 97 222 training images, 19 239 validation images, and 15 262 test images, with 2 174 unique CUIs. Synthetic captions were generated using an NXML-based context extraction workflow and a Qwen3-VL family vision-language model [30], followed by prompt-level hallucination control and manual review of sampled outputs. Concepts for the synthetic track were extracted with the same MedCAT-to-UMLS pipeline used for the main track, and source-linked metadata such as modality and IRMA labels were retained where available.

3.3 Participating Groups and Submitted Runs

The final test phase received 105 finished submissions across the five ImageCLEFmedical Caption subtasks. For the four automatically evaluated prediction leaderboards, a total of 135 finished runs were submitted: 54 for Main Concept Detection, 48 for Main Caption Prediction, 10 for Synthetic Concept Detection, and 23 for Synthetic Caption Prediction. The Explainability subtask received three manually evaluated submissions.

The Caption task received 16 Working Notes submissions from participating teams, of which 12 included open-source code describing their solutions.

3.4 Results

For Main Concept Detection, the best test score was achieved by `dsgt_caption`, reaching an F1-score of 0.5790, followed closely by `archimedes` with 0.5781. The third rank was shared by `UIT_Worker` and `basharderar`, both reaching an F1-score of 0.5725. For Main Caption Prediction, `AIstatLab` achieved the best overall score of 0.3755, followed by `UFPE-Essex-UNED` with 0.3603 and `dsgt_caption` with 0.3571. Across the main captioning leaderboard, relevance scores were consistently higher than factuality scores, indicating that factual precision remained a major limitation for generated medical captions.

On the synthetic tracks, `AUEB/DMST` achieved the best Synthetic Concept Detection score with an F1-score of 0.5609, followed by `krishnatewari` with 0.4741 and `NUSTPB` with 0.1048. Synthetic Caption Prediction produced the highest captioning scores of the benchmark, with `AIstatLab` reaching an overall score of 0.5840, followed by `AUEB/DMST` with 0.5136 and `UMUTeam` with 0.4632. Compared with the main captioning track, both relevance and factuality scores were

substantially higher on the synthetic-caption track, suggesting that synthetic captions formed a more regular and easier-to-model target distribution.

For the Explainability subtask, three submissions were manually assessed by a radiologist on a 1–5 scale. The best overall score was 3.69, achieved by a LLaVA-OneVision-Qwen2-based system using QLoRA/QDoRA. The manual review highlighted the importance of clinically plausible, readable captions as well as the need for explicit and well-aligned visual explanations.

3.5 Lessons Learned and Next Steps

The 2026 Caption task shows that medical concept detection has become increasingly competitive, with several teams achieving close scores on the main hidden-test leaderboard. Caption generation remains more difficult, especially with respect to factual correctness: systems often produce captions that are linguistically plausible and relevant, but still fail to capture clinically precise concepts. The difference between relevance and factuality scores suggests that future systems should focus more strongly on grounding, entity-level correctness, and clinically faithful generation.

The synthetic-data extension provided a useful stress test and complementary benchmark setting. Higher captioning scores on the synthetic track indicate that the synthetic captions are more standardized than real PubMed-derived figure captions, but they also provide a controlled setting for evaluating robustness to stylistic and lexical shifts. The explainability subtask showed that strong captions are often not sufficient.

4 The MEDIQA-CORE Task 1

Accurate brain tumor classification relies on integrating complementary information from MRI, histopathology, and clinical reports. In real-world clinical settings, however, modalities are often missing or acquired at different stages of the diagnostic workflow. MEDIQA-CORE Task 1 [5] addresses this gap by evaluating systems on their ability to effectively integrate heterogeneous data while remaining robust to missing modalities.

4.1 Task Setup

Accurate brain tumor classification relies on integrating complementary information from MRI, histopathology, and clinical reports. In real-world clinical settings, however, modalities are often missing or acquired at different stages of the diagnostic workflow. MEDIQA-CORE Task 1 addresses this gap by evaluating systems on their ability to effectively integrate heterogeneous data while remaining robust to missing modalities [5].

The challenge comprises three classification objectives: (1) Molecular subtype grouping into biologically meaningful tumor subtypes, (2) Lower-grade versus high-grade glioma (LGG vs. HGG) classification, (3)WHO histological grade prediction (Grades 1–2, 3, or 4).

4.2 Dataset

The task uses the CoRe-BT benchmark [18], a new multimodal glioma classification dataset comprising volumetric MRI, H&E-stained histopathology whole-

slide images (WSIs), and radiology reports collected from de-identified patients at a U.S. hospital. The dataset is designed to support tumor characterization under heterogeneous modality availability conditions and includes both raw data and precomputed embeddings to lower computational barriers for participation.

Provided modalities include multi-sequence MRI (T1, T1c, T2, and FLAIR), WSIs, radiology reports, and tumor segmentations. Participants were also given subject-level MRI embeddings extracted using NeuroVFM [25] and pathology tile- and slide-level embeddings generated using GigaPath [31]. Additional pathology preprocessing artifacts, including tissue segmentation previews and tile coordinate metadata, were provided to facilitate data exploration and quality assessment. By offering both raw multimodal data and foundation model-derived representations, the dataset enables participants to develop and evaluate multimodal learning approaches without requiring extensive image preprocessing or large-scale computational resources.

4.3 Results

Table 2 presents the official results on test set #1 under fully multimodal and missing-modality conditions. For comparison, we included our baseline system that employs a multimodal fusion framework combining text, MRI imaging, and WSI [18].

Table 2: Test-set #1 results under different modality-availability conditions.

Condition	Team	Level 1 F1	WHO F1	LGG/HGG F1	Avg.
Fully Multimodal	NeuroVietCH	1	0.956	1	0.985
	cnvntq	0.918	0.886	0.911	0.905
	UIT_Concabietbay	**0.878**	0.729	0.700	0.769
	DSGT	0.867	0.737	0.800	**0.801**
	Brihadaranyaka	0.540	0.772	**0.958**	0.757
	MEDIQA Organizers	*0.672*	***0.795***	*0.920*	*0.796*
MRI Reports Only	UIT_Concabietbay	0.018	0.267	0.438	0.241
	DSGT	0.330	0.464	0.556	0.450
	Brihadaranyaka	**0.823**	0.536	0.807	0.722
	MEDIQA Organizers	*0.589*	***0.726***	***0.920***	***0.745***
MRI Images Only	UIT_Concabietbay	0.666	0.611	0.807	0.695
	DSGT	0.472	0.579	**0.839**	0.630
	Brihadaranyaka	**0.844**	**0.625**	0.823	**0.764**
	MEDIQA Organizers	*0.284*	*0.242*	*0.438*	*0.321*
Pathology Images Only	UIT_Concabietbay	0.515	**0.756**	0.700	0.657
	DSGT	**0.557**	0.565	0.765	0.629
	Brihadaranyaka	0.540	0.696	**0.920**	**0.719**
	MEDIQA Organizers	*0.534*	*0.555*	*0.438*	*0.509*

On Test set #1, under the fully multimodal setting, `NeuroVietCH` achieved near-perfect performance across all tasks, obtaining an average F1 score of 0.985, while `cnvntq` ranked second with an average F1 score of 0.905. Among the teams with available modality-ablation results, `DSGT` achieved the highest average F1 score (0.801), slightly outperforming our baseline (0.796). `UIT_Concabietbay` obtained the best Level 1 F1 score (0.878), whereas `Brihadaranyaka` achieved the strongest LGG/HGG classification performance (0.958).

The modality-ablation results reveal distinct strengths across input modalities. In the MRI-reports-only setting, `Brihadaranyaka` achieved the strongest participant performance with an average F1 score of 0.722, although the baseline remained superior overall (0.745). For MRI-images-only evaluation, `Brihadaranyaka`

again obtained the highest average F1 score (0.764), while `DSGT` achieved the best LGG/HGG classification performance (0.839). In the pathology-images-only setting, `Brihadaranyaka` achieved the highest overall average F1 score (0.719), with `UIT_Concabietbay` obtaining the best WHO F1 score (0.756) and `DSGT` achieving the highest Level 1 F1 score (0.557). Overall, MRI images and pathology images appeared substantially more informative than MRI reports alone, as evidenced by the markedly higher performance achieved by most participating systems in the image-only settings.

4.4 Lessons Learned and Next Steps

The finalist teams explored a diverse range of multimodal learning strategies, including task-specific fusion, classical machine-learning ensembles, multimodal expert ensembles, and LLM-assisted reasoning. Although participants were provided with both the raw MRI and pathology data as well as the generated foundation-model embeddings derived from NeuroVFM and GigaPath, all finalist teams elected to incorporate the provided embeddings into their submissions. Detailed descriptions of the finalist systems and our baseline approach are presented in the MEDIQA-CORE Task 1 overview paper [5].

Several common trends emerged among the finalist solutions. All teams leveraged the provided foundation-model embeddings, underscoring the value of pretrained multimodal representations for computationally demanding pathology whole-slide images. Despite diverse fusion architectures, most top-performing approaches relied on ensembling across modalities, models, or cross-validation folds. Robust handling of missing data was key, using strategies such as missing-modality tokens, adaptive gating, imputation, and ensemble reweighting.

Future editions may further explore more challenging missing-modality scenarios, additional clinical data sources, and methods that better leverage foundation-model representations for multimodal clinical decision support.

5 The MEDIQA-CORE Task 2

Trainee-attending feedback and mentorship is a critical part of clinical training for radiologists. One principle medium of this training is an initial report generation task by a junior radiologist, followed by an attending radiologist providing corrections, clarifications, and comments in an addendum. In this task [6], we frame an edit-classification problem where suggested edits are given and acceptance status needs to be properly identified.

5.1 Task Setup

MEDIQA-CORE 2026 Task 2 [6] focuses on multimodal radiology report review and discrepancy assessment. Given a 3D abdominal CT examination, a preliminary radiology report, and a candidate report edit, participating systems must determine whether the proposed modification is clinically appropriate and assess the nature of the discrepancy. Systems are required to jointly reason over imaging and textual information in order to evaluate the edit across three classifications: (1)**Agreement Classification:** determine whether the proposed edit is supported by the CT findings. (Categories: Agree, Partially Agree, Disagree),

Table 3: MEDIQA-CORE 2026 Task 2: Official results on the Mixed Set.

Rank	Team	ID	Composite (Mixed)	Agreement (Mixed)	Severity (Mixed)	Edit Type (Mixed)
–	*MEDIQA Organizers*	797	0.39	0.68	0.56	0.82
1	Brihadaranyaka	1031	**0.33**	0.52	**0.52**	0.68
2	MediFact	1043	0.32	0.43	0.51	**0.70**
3	ClinicalVLM-TW	1037	0.22	**0.53**	0.38	0.63
4	IReL_IIT(BHU)	960	0.05	0.22	0.51	0.12

(2) **Severity Assessment:** estimate the clinical significance of the discrepancy introduced or corrected by the edit. (Categories: Critical, Moderate, Negligible), (3)**Edit Type Classification:** categorize the edit as a correction, addition, or clarification. (Categories: Correction, Addition, Clarification)

Evaluation was measured using accuracy per classification type as well as a composite score, which required matching all three classifications (with agreement classification collapsing "agree" and "partially agree" categories).

5.2 Dataset

The task used the RADAR dataset [29], a multimodal dataset for radiology report discrepancy analysis comprising abdomen and pelvis CT examinations from the ED at Harborview Medical Center (HMC), UW Medicine. Edits were generated using GPT-4o from real discrepancies between trainees and attending radiologists and reviewed by a board-certified radiologist. To combat data imbalance, plausible but incorrect edits were introduced as synthetic errors, also reviewed by a board-certified radiologist to eliminate implausible suggestions. The final benchmark covers 50 abdominal CT examinations (20 validation, 30 test), comprising 319 image series and 45,838 slices. The test set contained 153 natural candidate edits and 274 mixed edits augmented with expert-verified synthetic "Disagree" edits.

5.3 Results

Among the four finalist teams, the `Brihadaranyaka` team employed a text-only approach using a single inference call to Claude Sonnet 4 to predict all three categories. Two teams (`MediFact and IReL_IIT(BHU)`) opted to use a text-embedding model and a 3D CNN for encoding images before adding feature-fusion and decision layers. The former team used a multi-task objective function whereas the latter team adopted threshold-based layers. The latter team also used data augmentation to introduce more training data. The `ClinicalVLM-TW` team featured multiple step modules, including multi-slice reading and selection, as well as additional injection of evidence including calls to OpenEvidence. They also used an ensemble of Claude Opus, Claude Sonnet, GPT-4o, and Gemini 2.5 with majority voting.

5.4 Lessons Learned and Next Steps

Detailed descriptions of the finalist systems and our baseline approach are presented in the MEDIQA-CORE Task 2 overview paper [6]. We observed that carefully crafted clinically aligned prompts can perform well, and that text embeddings with feature fusion can compete with large language models. A complex

multi-step ensemble approach did not yield immediate performance boosts over simpler methods.

We envision this work will be directly relevant as AI-assisted radiology report generation and editing services become increasingly commonplace, and further studies on verification, quality assurance, and discrepancy analysis become more critical.

6 The GANs Task

Building on the previous three editions [3, 4, 2], the 2026 edition expands the scope of the challenge by introducing two new subtasks alongside the established fingerprint detection problem, addressing privacy leakage from complementary angles: the reversibility of the latent-to-image mapping and the trade-off between image fidelity and patient privacy in generative models. The task is described in detail in the overview paper [1].

6.1 Task Setup

The 2026 edition comprises three subtasks that address distinct but interrelated aspects of privacy in synthetic medical image generation.

Subtask 1. Detect Training Data Usage. Continuing from all previous editions, participants must determine for each real image in the test set whether it was used to train the GAN that produced a given set of synthetic outputs. The task is formulated as binary classification (label 1 = used, label 0 = not used), and is evaluated primarily using Cohen's Kappa, with additional metrics including accuracy, sensitivity, specificity and F1 score.

Subtask 2.Identify Corresponding Latent Vectors. This newly introduced subtask evaluates how faithfully generative models map latent vectors to output images and whether this mapping reveals structural or patient-specific information. Participants are given 10,000 synthetic images generated by a BigGAN-Deep model alongside an unordered pool of their corresponding 128-dimensional latent vectors sampled from $\mathcal{N}(0, I_{128})$, and must establish the correct one-to-one correspondence. Performance is measured by matching accuracy: the proportion of image-latent pairs for which the predicted index matches the ground truth.

Subtask 3.Privacy-Preserving CT Slice Generation. Participants are asked to train a generative model on a provided set of real CT slices and submit 5,000 synthetic images together with the generator checkpoint. Submitted images are evaluated across two dimensions: image quality, measured by the Fréchet Inception Distance (FID) computed with a domain-specific feature extractor; and privacy preservation, quantified through a composite Privacy Preservation Score ($\text{PPS} = 1 - \text{mean}(\ell_1, \ell_2, \ell_3, \ell_4)$) derived from four independent privacy attacks: nearest-neighbour distance audit (ℓ_1), membership inference (ℓ_2), patient re-identification (ℓ_3), and white-box inversion (ℓ_4). All attacks use a held-out patient set, consisting of CT slices from patients never released to participants.

6.2 Data Set

The benchmarking dataset consists of axial slices of 3D CT scans from approximately 8,000 lung tuberculosis patients, stored as 256×256 8-bit grayscale PNG

images. For Subtask 1, the development set contains 10,000 GAN-generated synthetic images and 6,400 annotated real images (3,200 used for training the GAN, 3,200 not used); the test set contains 2,140 real images with unknown labels. For Subtask 2, the development set consists of 10,000 image-latent vector pairs generated by BigGAN-Deep; the test set provides an additional 10,000 images with their latent vectors shuffled. For Subtask 3, participants are provided with 10,000 unlabelled CT slices for model training, drawn from the same patient cohort.

6.3 Results

Table 4 provides a summary of the best submission per team across all subtasks.

Table 4: Summary of best results per team across all three GANs subtasks.

Team	**Subtask 1**	**Subtask 2**	**Subtask 3**	
	Cohen's κ ↑	Matching Acc. ↑	FID ↓	PPS ↑
Bharathi & Harish	-0.0383	–	–	–
csmorgan	-0.0290	0.0224	3.2358	0.9086
CVG-IBA	-0.0243	0.3677	–	–
DS@GT ARC	–	–	0.3290	0.5490
Agentili	–	1.0000	–	–
AI Multimedia Lab*	-0.0056	0.3737	1.2082	0.6041

For Subtask 1, all submitted runs achieved Cohen's κ scores at or below zero, ranging from -0.0804 to -0.0056, consistent with previous editions. The best-performing team was `CVG-IBA` ($\kappa = -0.0243$, accuracy=0.4879), followed by `csmorgan` ($\kappa = -0.0290$, accuracy=0.4855). These results suggest that the studied GAN does not leave easily detectable training-data fingerprints in its outputs, which represents an encouraging outcome from a privacy standpoint, while simultaneously highlighting the limitations of current detection frameworks.

For Subtask 2, results spanned a wide range, from near-random performance (0.0003, `csmorgan`) to a perfect matching accuracy of 1.0000 (`Agentili`). The `Agentili` team demonstrated that supervised image-to-latent regression with globally consistent one-to-one assignment via the Hungarian algorithm is sufficient to recover all 10,000 image-latent correspondences without error. This has direct privacy implications: an adversary with access to generated images can recover the originating latent codes with high accuracy.

For Subtask 3, a clear privacy-utility trade-off emerged, with no team achieving simultaneously strong privacy preservation and high image fidelity. The csmorgan team achieved the highest PPS (0.9086) using a compact DCGAN-style generator with strong R_1 regularisation and a weak discriminator, at the cost of substantially degraded image quality (FID = 3.2358). The DS@GT ARC team explored ten configurations combining a Differentially Private OT-CFM generator with a post-generation geometric filtering pipeline; their best PPS was 0.549 and best FID was 0.2639. The AI Multimedia Lab's StyleGAN2 submission serves as a baseline from the task organizing team. Across all submissions, patient re-identification (ℓ_3) emerged as the dominant leakage vector, with all high-fidelity submissions scoring above 0.90 regardless of the privacy mechanism employed, suggesting that patient-level identity cannot be removed by output-level filtering alone.

6.4 Lessons Learned and Next Steps

The 2026 edition provided three key insights. First, image-level membership inference against well-regularised GANs remains extremely difficult under black-box conditions, consistent with previous editions. Second, the BigGAN-Deep latent-to-image mapping is recoverable with perfect accuracy using supervised regression, with direct privacy implications for latent-space security. Third, patient re-identification emerged as the most persistent leakage vector in synthetic CT generation, resisting both differential privacy training and post-hoc geometric filtering. Future editions will refine the privacy evaluation protocol, explore new generation architectures, and introduce tasks further probing the boundaries of privacy, utility, and generalization in generative medical imaging.

7 The MedVQA-GI Task

The MedVQA-GI task is the fourth edition of the Medical Visual Question Answering (VQA) for the Gastrointestinal (GI) Tract challenge at ImageCLEF [19, 20, 12]. Building on the previous three editions, the 2026 edition shifts its focus from data and image generation toward trustworthy multimodal reasoning in realistic clinical settings. Participants must not only answer clinically relevant questions about GI endoscopy images, but also justify their answers in a manner aligned with medical reasoning while adhering to clinical safety principles, explicitly addressing emerging concerns around overconfidence, hallucination, and misleading explanations in medical artificial intelligence systems.

7.1 Task Setup

The 2026 edition comprises two tasks that share the same images and question pool but differ in their evaluation focus.

Task 1 (Clinically Relevant VQA). Participants develop algorithms that answer clinically relevant questions about GI endoscopy images, correctly interpreting both the visual findings and the intent of the question to produce concise, medically sound answers, with a focus on the diagnostic and descriptive questions commonly encountered in clinical practice. Submissions are scored with automatic text-generation metrics that capture both answer correctness and language quality, BLEU, ROUGE-1/2/L and METEOR.

Task 2 (Multimodal Explainability and Safety-Aware Reasoning). This task extends standard VQA by requiring models to produce coherent multimodal justifications that combine a clinician-oriented textual explanation with visual evidence drawn from the image, such as heatmaps, segmentation masks, or bounding boxes, optionally accompanied by a confidence score. Beyond answer correctness, submissions are judged by expert-based assessment of explanation quality, interpretability, and clinical plausibility, together with a dedicated behavioral-safety evaluation that probes for undesirable behaviors such as overconfidence, hallucination, and unsafe clinical recommendations. Task 1 is therefore evaluated fully automatically, whereas Task 2 combines automatic measures with expert and safety assessment.

Table 5: Task 1 automatic evaluation results for all seven participating teams, on the public Kvasir-VQA-test set and the held-out private test set. Metrics are BLEU-1/2 (B-1/B-2), ROUGE-1/2/L (R-1/R-2/R-L), and METEOR (MET). Rows are ordered by METEOR on the public set, and the best value in each column is shown in **bold**. Strong scores on the public set largely fail to transfer to the unseen private set, and the ranking essentially inverts.

	Kvasir-VQA-test (public)						Private (held-out)					
Team	B-1	B-2	R-1	R-2	R-L	MET	B-1	B-2	R-1	R-2	R-L	MET
GutCheck	**0.748**	**0.635**	**0.893**	**0.133**	**0.888**	**0.498**	0.001	0.000	0.107	0.013	0.105	0.038
TUW_AIR22	0.730	0.560	0.840	0.090	0.820	0.460	0.000	0.000	0.130	0.010	0.130	0.040
UIT_NEWRON	0.640	0.547	0.802	0.091	0.800	0.435	0.001	0.000	0.137	0.007	0.135	0.047
IReL_IIT(BHU)	0.175	0.100	0.601	0.058	0.599	0.374	0.250	0.200	0.390	0.250	0.350	0.480
OmniMed	0.255	0.191	0.664	0.029	0.664	0.327	0.000	0.000	0.138	0.001	0.138	0.044
CSMorgan-MEDVQA	0.192	0.119	0.516	0.010	0.505	0.259	0.004	0.002	0.199	0.023	0.192	0.086
Something Creative	0.030	0.000	0.050	0.000	0.050	0.050	**0.520**	**0.400**	**0.570**	**0.360**	**0.530**	**0.520**

7.2 Data Set

The task uses Kvasir-VQA-x1 [13], an expanded version of Kvasir-VQA [14] that draws on publicly available GI endoscopy collections, including HyperKvasir [7], Kvasir-SEG [24], and Kvasir-Instrument [23]. Kvasir-VQA-x1 comprises more than 150,000 question-answer pairs derived from GI endoscopy images spanning a wide range of anatomical sites and pathological findings, and supports both answer prediction and explainability evaluation. The questions are clinically motivated and diverse, ranging from diagnostic and descriptive queries to questions requiring classification, reasoning, spatial localization, and attribute recognition. The development split is publicly available.[19]

As in previous editions, the final evaluation uses a separate, held-out private test set sampled from images not included in the development data. This set is distributionally distinct and unseen during development, which lets us assess how well participating systems generalize to genuinely novel GI images rather than to the public split alone.

7.3 Results

Table 5 reports the Task 1 automatic metrics for all seven teams on both the public Kvasir-VQA-test set and the held-out private test set. The two sets tell very different stories. On the public set, GutCheck, TUW_AIR22, and UIT_NEWRON achieve strong scores (BLEU-1 of 0.748, 0.730, and 0.640, respectively, and METEOR up to 0.498), yet on the private set the same systems collapse to near-zero BLEU and METEOR below 0.05. Conversely, the two teams with comparatively modest public scores, IReL_IIT(BHU) and Something Creative, are the only systems whose performance holds up on the unseen data, reaching the highest private METEOR (0.480 and 0.520) and ROUGE-L (0.350 and 0.530).

7.4 Lessons Learned and Next Steps

The dominant finding of this edition is that strong in-distribution performance did not transfer to the held-out private set. Several systems that topped the

[19] `https://datasets.simula.no/kvasir-vqa-x1`

public Kvasir-VQA-test leaderboard scored close to zero on the unseen private images, which points to overfitting to the public split and sensitivity to distribution shift rather than to robust clinical reasoning. This confirms the value of evaluating on a separate, unreleased test set: public leaderboard scores alone substantially overstate how well current VQA systems generalize. Robustness and generalization to unseen GI data therefore stand out as the central open problems for the field.

These observations also motivate the explainability and safety focus of Task 2: models that answer well in-distribution but cannot justify their answers, or that remain overconfident when out of distribution, are of limited clinical value. As next steps, the expert-based and behavioral-safety results for Task 2 will be presented at the workshop, and future editions will expand the diversity of the held-out evaluation data and further develop the trustworthiness and safety dimensions of the assessment.

8 The Deepfake Task

Deepfakes are AI-generated media that portrays real individuals and often times aims to mislead or misinform the public. With generative models being more widely available then ever, many researchers aim to study ways to protect against the threat of deepfakes in society. The Deepfake task, in its first edition, aims to bridge the gap between real-life model performance and experimental results in Deepfake Detection, with a particular emphasis on the generation process.

8.1 Task Setup

The task is held on two different data modalities: **Images** and **Audio** and is organized into two subsets which are interconnected:

- **Deepfake Generation**: Given a set of real video or audio files, participants are asked to generate deepfakes under specific constraints, including identity, realism, and content requirements (e.g., facial landmarks for images and a prescribed phrase for audio). The resulting deepfakes are then evaluated in the second subtask using detection models, with performance measured in terms of evading detection.
- **Deepfake Detection**: Given a dataset comprising real samples, organizer-generated deepfakes, and participant-generated deepfakes, participants are required to develop models that accurately assess the authenticity of the provided data.

8.2 Data Set

For the **Generation** task, participants were provided with real data spanning multiple identities:

- **Images** Participants received 335 development videos (15–40 seconds each), each corresponding to a different identity, extracted from CC-BY license Youtube Videos. For generation, facial landmark constraints were provided as targets for the synthesized images. These landmarks were extracted from three real secret frames per identity.

– **Audio** Participants were provided with 16 speakers (11 male, 5 female), aged 22–32, recorded by the organizers. All recordings were 16-bit monophonic .wav files in English. The data were divided into three text-length categories (Short, Medium, Long), each containing at least five audio–text pairs, resulting in 30 reference files per speaker. For generation, participants were given speaker identities and target phrases to be spoken by the corresponding identities.

For the **Detection** task, participants received a binary classification dataset comprising: (i) real data from the generation reference set, (ii) real data from a secret set, (iii) organizer-generated deepfakes, and (iv) participant-generated deepfakes. In total: 24,404 images (20,002 deepfakes, 4,002 real) and 11,520 audio files (10,560 deepfakes, 960 real).

8.3 Results

The task, including data, evaluation procedure, and results is described in detail in our overview paper [28].

For **Deepfake Generation**, teams were evaluated on a Quality score (constraint adherence and identity similarity) and an Evasion score (weighted average evasion against participant and organizer detectors), with the final score as their product. For **Audio**, Quality ranged from 0.49–0.73 and Evasion from 0.28–0.59. Team `Go-To-Germany` achieved the best scores (Quality: 0.734, Evasion: 0.51) using four Text-To-Speech models across identities. For **Images**, Quality ranged from 0–0.70 and Evasion from 0.28–0.59. Team `Brihadaranyaka` achieved the highest overall score (Quality: 0.70, Evasion: 0.62) with a landmark-constrained adversarial generation pipeline. Team `Go-To-Germany` finished second, with stronger evasion but weaker quality.

For **Deepfake Detection**, the score was a weighted average of accuracy across the four data types. For **Audio**, team `Go-To-Germany` achieved the best result with a weighted accuracy of 0.95 using a four-backbone self-supervised learning ensemble. For **Images**, teams `k20221289` and `Go-To-Germany` scored 0.6936 and 0.6987 respectively, both using multi-backbone ensemble detectors with fusion approaches.

8.4 Lessons Learned and Next Steps

Overall, the first edition of the task was successful and provided valuable insights into deepfake detection. In audio deepfake detection, several teams proposed innovative approaches that achieved strong generalization. In contrast, image deepfake detection results were often closer to random, indicating that generalization in this modality remains challenging. For the generation task, a weak correlation was observed between perceptual quality and evasion score, suggesting that, in some cases, producing visually realistic deepfakes may be sufficient to deceive detectors.

9 The ToPicto Task

The ImageCLEFtoPicto task, in its third edition, continues to advance research on the automatic conversion of text into meaningful sequences of pictograms de-

signed to support Alternative and Augmentative Communication (AAC). AAC encompasses a set of strategies and methods used to supplement or compensate for oral language, helping individuals with communication disorders convey their thoughts, needs, and emotions. By providing tools that automatically translate text into pictograms, the task aims to reduce communication barriers between oralizing individuals with no prior knowledge of pictograms and AAC users.

9.1 Task Setup

In summary, the two tasks are:

- **Text-to-Picto** extends the previous editions while introducing a major shift: the source language is English rather than French. Participants are required to generate coherent sequences of pictogram labels from English text, adapting their systems to account for the linguistic, syntactic, and semantic differences involved. Each pictogram term is linked to a unique ARASAAC pictogram. The problem is framed as a translation task, where the source is English text and the target is a sequence of pictogram terms.
- **Next-Pictogram Prediction**, is a newly introduced subtask inspired by Pereira et al. [26]. It focuses on predicting the most probable next pictogram in a given sequence, with performance evaluated using a top-N accuracy metric.

9.2 Data Set

The dataset for Task 1 consists of aligned pairs of English text and ARASAAC pictograms. Aragonese Centre for Augmentative and Alternative Communication (ARASAAC) is an online resource of more than 25,000 pictograms distributed under a Creative Commons BY-NC-SA license, funded by the Department of Culture, Sports, and Education of the Government of Aragón (Spain). The English resource was constructed by translating a subset of the existing ProPicto-Orféo corpus into English (approx. 63,000 sentences). As in previous editions, the data is provided in JSON format, with each entry containing a unique utterance identifier, the source text, the target sequence of pictogram terms, and the corresponding list of ARASAAC pictogram identifiers. For Task 2, uses the same underlying data as the Text-to-Picto subtask. For training, the same approximately 63,000 sentences are provided. At test time, the model is given all but the last token of a pictogram sequence and must predict the missing final token. The test file contains 1,986 instances. To prevent participants from recovering the withheld final token by matching it against the source corpus, the id field in the Next-Pictogram Prediction data serves as a surrogate identifier rather than the original CommonVoice identifier.

9.3 Results

The results for the ToPicto subtasks are presented in Table 6 and Table 7. All scores are reported as percentages, expressed on a 0–100 scale. For Text-to-Picto, the top-ranking teams achieved remarkably close performances, separated by only decimals. The considerably lower scores in Next-Pictogram Prediction reflect the intrinsic difficulty of the task, owing to the scarcity of training data

relative to mainstream language modeling tasks. The results invite an interesting comparison with next-word prediction systems for languages such as English: the considerably lower scores reflect the intrinsic difficulty of the task, owing in part to the scarcity of training data relative to mainstream language modeling tasks.

Table 6: ToPicto 2026 Text-to-Picto Translation Results, best run per team, ordered by PictoER (lower is better).

Rank	Participant	sacreBLEU ↑	METEOR ↑	PictoER ↓
1	psp075	52.55	71.69	31.48
2	jlcuad	52.00	71.02	31.99
3	gokul_n_v	52.21	71.04	32.12
4	kaiyili	51.34	70.40	32.55
5	fuchuanye	51.34	70.40	32.55
6	sami_haq	51.33	70.36	32.75
7	hiwiy	50.38	69.67	33.33
8	melody	48.49	69.98	35.62
9	lekshmivitscope	28.76	63.92	38.84
10	varghesekjames	31.23	61.18	53.77
11	linxiaocan	22.73	48.59	57.79

Table 7: ToPicto 2026 Next-Pictogram Prediction Results, best run per team, ordered by Top-1 accuracy (higher is better).

Rank	Participant	Top-1 ↑	Top-3 ↑	Top-5 ↑	Top-10 ↑	Sem. Sim. ↑
1	melody	8.91	14.65	18.08	22.21	34.38
2	fuchuanye	7.85	12.84	14.80	19.34	32.36
3	hiwiy	7.80	12.39	14.15	16.97	33.37
4	kaiyili	7.25	12.49	15.21	19.74	31.94
5	linxiaocan	6.50	12.64	16.11	21.80	32.50
6	tce_chidambaram	5.44	10.42	14.30	19.34	32.41
7	saisaranya	5.29	10.27	13.39	16.41	32.46
8	sami_haq	3.78	9.47	12.89	19.89	27.02

10 The Multimodal Reasoning Task

The ImageCLEF 2026 Multimodal Reasoning task aims to benchmark multimodal reasoning systems on challenging vision-based academic examination questions, covering both multiple-choice and open-ended settings. The task is designed to assess the ability of vision-language models to understand question images, reason over visual and textual content, and produce accurate answers under realistic exam-like conditions. To promote open-source and open-weight development within the public research community, this year the task is restricted to open-weight or open-source models.

10.1 Task Setup

The benchmark is organized into two complementary subtasks:

- **Multiple-Choice Question Answering (MCQ)** focuses on questions provided as images together with three to five candidate answers. Systems are required to process the visual question, reason over the available information, and select exactly one correct option.
- **Open Question Answering (OpenQA)** addresses a more unconstrained setting in which the question is provided as an image but no answer candidates are given. Systems must generate a free-form textual response, making

this subtask a direct test of both multimodal understanding and generative reasoning.

Both tasks have two separate leaderboards, for questions with and without visual elements.

10.2 Data Set

The dataset was collected from publicly available examination materials ranging from primary to high-school levels, with most questions targeting the high-school level. Due to the limited availability of open-ended questions in some languages, the OpenQA subset was supplemented with primary- and middle-school Olympiad materials. In addition, some open-ended questions were created by repurposing MCQ-style questions from Exams-V [10]: the answer choices were removed, and the full text of the correct answer, rather than the option letter, was used as the gold label. The subject coverage is diverse, with the largest portions coming from Mathematics, Physics, Chemistry, Geography, and Biology, and additional questions from Informatics, History, Psychology, Sociology, Philosophy, Fine Arts, Music, and other social science and humanities domains. Table 8 presents the dataset distribution.

Table 8: MultimodalReasoning dataset distribution by task type, presence of visual elements, and language.

Subset	Bulgarian	Chinese	Croatian	English	Italian	Serbian	Total
OpenQA Visual	209	264	221	191	181	141	1,207
OpenQA Non-visual	200	149	200	200	150	150	1,049
MCQ Visual	110	342	57	500	54	54	1,117
MCQ Non-visual	451	267	335	240	173	187	1,653

10.3 Results

The results show that multimodal academic question answering remains challenging in the open-ended setting. In the MCQ track, performance was substantially higher than in OpenQA, reflecting the advantage of having answer candidates available in the image. For questions with visual elements, spirosbax achieved the best MCQ score with an overall score of 0.8406, followed by DS@GT with 0.7986 and FAU with 0.7108. For MCQ questions without visual elements, pratikpriyanshu ranked first with 0.7538, followed by linxiaocan with 0.7084 and rafiulbiswas with 0.6152.

The OpenQA results were lower overall, confirming that direct answer generation is a substantially more difficult setting than answer selection. For OpenQA questions with visual elements, FAU obtained the highest COMET score of 0.6488, followed by wangshou66 with 0.6366 and uned-martinez with 0.5938. For OpenQA questions without visual elements, linxiaocan achieved the best score with 0.6563, followed by pratikpriyanshu with 0.5285. The relatively modest scores in the OpenQA track suggest that current open-weight systems still struggle to combine visual interpretation, multilingual question understanding, and precise answer generation in realistic exam-style settings.

Table 9: Overall leaderboard results for MCQ and OpenQA subtasks.

MCQ

With visual elements

Rank	Participant	Overall
1	spirosbax	0.8406
2	DS@GT	0.7986
3	FAU	0.7108
4	zhaijinghe	0.6150
5	sjaini	0.5927
6	begyed	0.5774
7	linxiaocan	0.5560
8	pratikpriyanshu	0.5076
9	wether	0.4673

Without visual elements

Rank	Participant	Overall
1	pratikpriyanshu	0.7538
2	linxiaocan	0.7084
3	rafiulbiswas	0.6152

OpenQA

With visual elements

Rank	Team	COMET
1	FAU	0.6488
2	wangshou66	0.6366
3	uned-martinez	0.5938
4	liuzhe3085	0.5829
5	wenlong	0.5497
6	rafiulbiswas	0.5143
7	linxiaocan	0.5117
8	pratikpriyanshu	0.4286

Without visual elements

Rank	Team	COMET
1	linxiaocan	0.6563
2	pratikpriyanshu	0.5285

10.4 Lessons Learned and Next Steps

The gap between MCQ and OpenQA performance confirms that multiple-choice evaluation alone can overestimate model capabilities, as answer candidates provide strong cues that are absent in free-form generation. Open-weight vision-language models remain limited on multilingual open-ended educational QA, particularly when questions involve diagrams, graphs, or tables that require connecting visual extraction with subject-specific reasoning. More, evaluation remains an important limitation. Metrics such as COMET and lexical-overlap measures are useful for scalable ranking, but they do not always capture semantic correctness, especially for short answers, numerical answers, equivalent mathematical forms, or valid paraphrases.

11 The Ai4Agri Task

11.1 Task Setup

The Ai4Agri task is organized into two sub-tasks targeting distinct agricultural remote sensing problems using on satellite imagery.

Subtask 1 – AgriPotential: Viticulture Edition challenges participants to predict the agricultural potential of land for viticulture (grapevine cultivation) using high-resolution, multi-temporal remote sensing data. Agricultural potential is defined by a set of environmental factors (soil, climate, topography) that determine the suitability for a specific crop. Using multi-spectral and multi-temporal Sentinel-2 satellite imagery from 2017 to 2019, participants must build a model capable of predicting the five-class ordinal potential of each land pixel: very low, low, average, high, and very high.

Subtask 2 – DACIA5 aims to advance agricultural practices through remote sensing, focusing on innovative uses of the DACIA5 [22] dataset. It comprises two sub-challenges: (1) *Crop Identification: Past vs. Present*: models are trained on data from 2020–2023 and evaluated on unseen 2024 data; and (2) *Early Crop Identification*: targeting early identification of winter wheat and al-

falfa in March, where excluded March patches from 2022–2024 form the test set.

11.2 Dataset

Subtask 1 uses a subset of the AgriPotential benchmark [11], covering diverse agricultural landscapes in Southern France. The dataset contains 34 Sentinel-2 images collected between 2017 and 2019, each with 10 spectral bands acquired at irregular time intervals. Ground truth consists of pixel-level annotations representing viticulture potential across five ordinal classes.

Subtask 2 uses the DACIA5 dataset [22], which comprises all available Sentinel-1 SAR and Sentinel-2 MSI imagery from 2020–2024, covering an agricultural area north of Brașov city, Romania. The dataset includes Sentinel-2 images ($800 \times 450 \times 12$ spectral bands) and Sentinel-1 images ($800 \times 450 \times 2$ radar channels) in GeoTIFF format, along with 32×32 pixel optical and radar patches. It features 17 distinct crop types, with pixel-level masks validated by the *National Institute of Research and Development for Potato and Sugar Beet* (NIRDPSB).

11.3 Results

Subtask 1. The final leaderboard was led by team `snspace` with an Accuracy±1 score of 69.08%, followed by `DS@GT AI4Agri` (68.32%), `ronny_toro` (65.66%), and `titouann` (65.38%). The remaining teams, `janani_sp` and `vb7132`, obtained scores of 56.15% and 51.28%, respectively. Most participants initially adopted standard U-Net-style segmentation architectures; approaches later diversified toward pretrained models and alternative formulations of the ordinal prediction task. Our internal benchmark exceeded 80% Accuracy±1, indicating substantial remaining headroom.

Subtask 2. A single submission was received from team `Innovware`, who developed a CropCNN architecture for classifying Sentinel-2 multispectral patches into seven crop categories (Wheat, Corn, Peas, Rapeseed, Potato, Sugarbeet, Alfalfa). On the 2024 test set, the model achieved an Overall Accuracy (OA) of 38.83% and an Average Accuracy (AA) of 26.88%, resulting in a Q_1 Score of 32.86.

11.4 Lessons Learned and Next Steps

Both sub-tasks revealed a gap between participant submissions and attainable performance, pointing to shared challenges in generalization and task complexity.

For Subtask 1, the primary finding is that while participants identified strong building blocks, a visible gap remains between the best participant result (69.08%) and the internal benchmark (>80%). From an organizational standpoint, determining how to properly conceal the test set while defining a clear submission protocol proved non-trivial. Future editions will focus on consolidating the competition infrastructure, particularly around test set management and submission format standardization.

For Subtask 2, the key lesson is that a lightweight CNN architecture is a reasonable baseline, but performance was significantly impacted by the temporal domain shift between training and test years, severe class imbalance for minority

crops (e.g., Potato and Sugarbeet), and frequent misclassification of spectrally similar crops (e.g., Wheat and Corn). Future iterations should explore architectures with stronger temporal generalization and strategies to address class imbalance more effectively.

12 Conclusion

This paper presents an overview of the ImageCLEF 2026 benchmarking campaign. We introduce five main tasks addressing challenges in medicine, classification, retrieval, question answering, generation, security and related areas.

Most approaches were based on deep learning and convolutional neural networks. In many cases, the best performance on the proposed tasks was achieved by combining multiple complementary models through fusion strategies. Additionally, large language models (LLMs) and vision-language models (VLMs) have increasingly gained popularity as general-purpose approaches applicable to a wide range of tasks. A recurring theme across tasks was the challenge of generalization – whether across modalities, time, or unseen data distributions – pointing to an area where significant progress remains to be made.

ImageCLEF continues to serve as a platform for innovation, learning, and progress across diverse fields. Future editions will focus on improving existing tasks, expanding into new domains, increasing participation, and encouraging experimentation in emerging areas, while addressing current limitations.The overall goal is to continuously enhance the quality of the benchmarking campaign and support meaningful advances in the field.

Acknowledgements

The work of Benjamin Bracke, Lea Reinartz and Raphael Brüngel was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Henning Schäfer, Tabea M. G. Pakull, Hendrik Damm, Praveen Nath, and Bahadır Eryılmaz was funded by a PhD grant from the DFG Research Training Group 2535 Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed). The work of Dimitar Dimitrov, Ivan Koychev, Dimitrina Zlatkova, and Georgi Pachov is partially supported by the EU NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008 and UNITe BG16RFPR002-1.014-0004 funded by PRIDST. The work of Alexandra-Georgiana Andrei, Dan-Cristian Stanciu was funded by the Ministry of Research, Innovation and Digitization, UE-FISCDI, Solutions Axis, grant SOL4/2024, project number PN-IV-P6-6.3-SOL-2024-2-0320. The work of Mihai Gabriel Constantin, Dan-Cristian Stanciu was supported by UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-0060, DeteRel, 1Sol(T12)/2024.

The toPicto task was co-funded by the European Union under the Marie Skłodowska-Curie Grant Agreement No 101081465 (AUFRANDE). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the Research Executive Agency, which cannot be held responsible for them. This project was also funded by the Agence Nationale de la Recherche (ANR) through the project PANTAGRUEL (ANR-23-IAS1-0001). This work is also carried out as part of the AugmentIA Chair, hosted by the Grenoble INP Foundation with sponsorship from the Artelia Group, and supported by the National Research Agency (ANR), under the France 2030 program with reference ANR-23-IACL-0006 (MIAI Cluster). The pictographic symbols used are the property of the Government of Aragón and have been created by Sergio Palao for ARASAAC (`http://www.arasaac.org`), which distributes them under Creative Commons License BY-NC-SA.

The DACIA5 and AgriPotential Subtasks, forming the broader Ai4Agri Task, were funded by the AI4AGRI project, officially titled "Romanian Excellence Center on Artificial Intelligence on Earth Observation Data for Agriculture". The AI4AGRI project received funding from the European Union's Horizon Europe research and innovation program under grant agreement no. 101079136.

Declaration on Generative AI

During the preparation of this work, the authors used generative models (Claude Anthropic) in order to: enhance grammar, spellcheck and paraphrase. All scientific content and references were provided by the authors. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

1. Andrei, A., Constantin, M., Dogariu, M., Karpenka, D., Prokopchuk, Y., Radzhabov, A., Stanciu, D., Ştefan, L., Kovalev, V., Müller, H., Ionescu, B.: Overview of the 2026 ImageCLEFmedical GANs task: Fingerprint detection, latent space analysis, and privacy-preserving medical image synthesis. In: CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany (September 21–24 2026)
2. Andrei, A.G., Constantin, M.G., Dogariu, M., Radzhabov, A., Ştefan, L.D., Prokopchuk, Y., Kovalev, V., Müller, H., Ionescu, B.: Overview of ImageCLEFMedical 2025 GANs task: Training data analysis and fingerprint detection. In: CLEF2025 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain (September 9-12 2025)
3. Andrei, A., Radzhabov, A., Coman, I., Kovalev, V., Ionescu, B., Müller, H.: Overview of ImageCLEFmedical GANs 2023 task – Identifying Training Data "Fingerprints" in Synthetic Biomedical Images Generated by GANs for Medical Image Security. In: CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
4. Andrei, A., Radzhabov, A., Karpenka, D., Prokopchuk, Y., Kovalev, V., Ionescu, B., Müller, H.: Overview of 2024 ImageCLEFmedical GANs Task – Investigating Generative Models' Impact on Biomedical Synthetic Images. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 9-12 2024)
5. Ben Abacha, A., Heras Rivera, J.E., Low, D.K., Yim, W., Ruzevick, J., Child, D., Kurt, M.: Overview of the mediqa-core 2026 task 1 on brain tumor subtype classification. In: CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany (September 2026)
6. Ben Abacha, A., Sun, Z., Yim, W., Xia, F., Yetisgen, M.: Overview of the mediqa-core 2026 task 2 on radiology report discrepancy assessment. In: CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany (September 2026)
7. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1) (2020). https://doi.org/10.1038/s41597-020-00622-y
8. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003) (2004)
9. Damm, H., Pakull, T.M.G., Reinartz, L., Bracke, B., Eryılmaz, B., Nath, P., Brüngel, R., Schmidt, C.S., Schäfer, H., Ben Abacha, A., García Seco de Herrera, A., Müller, H., Friedrich, C.M.: Overview of ImageCLEFmedical 2026 – medical concept detection and caption generation with synthetic data extensions. In: CLEF2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany (September 21-24 2026), to appear
10. Das, R., Hristov, S., Li, H., Dimitrov, D., Koychev, I., Nakov, P.: EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision

language models. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7768–7791. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.420
11. El Sakka, M., De Pourtales, C., Chaari, L., Mothe, J.: Agripotential: A novel multi-spectral and multi-temporal remote sensing dataset for agricultural potentials. In: 2025 International Conference on Content-Based Multimedia Indexing (CBMI). pp. 1–6. IEEE (2025)
12. Gautam, S., Halvorsen, P., Riegler, M.A., Thambawita, V., Hicks, S.A.: Overview of ImageCLEFmedical 2025 – medical visual question answering for gastrointestinal tract. In: CLEF2025 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain (September 9-12 2025)
13. Gautam, S., Riegler, M.A., Halvorsen, P.: Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy. arXiv preprint arXiv:2506.09958 (Jun 2025). https://doi.org/10.48550/arXiv.2506.09958
14. Gautam, S., Storås, A., Midoglu, C., Hicks, S.A., Thambawita, V., Halvorsen, P., Riegler, M.A.: Kvasir-VQA: A text-image pair gi tract dataset. In: Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications (VLM4Bio). pp. 3–12. ACM (2024). https://doi.org/10.1145/3689096.3689458
15. Gautam, S., Thambawita, V., Riegler, M.A., Halvorsen, P., Hicks, S.A.: Overview of ImageCLEFmedical 2026 - medical visual question answering for gastrointestinal tract. In: CLEF2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany (September 2026)
16. Hagen, M., Potthast, M., Stein, B., Schaer, P., Zangerle, E., MacAvaney, S., Struß, J.M., Salido, E.S., Barrón-Cedeño, A., de Herrera, A.G.S. (eds.): Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026). Lecture Notes in Computer Science (LNCS), Springer (2026)
17. Hagen, M., Potthast, M., Stein, B., Schaer, P., Zangerle, E., MacAvaney, S., Struß, J.M., Salido, E.S., de Herrera, A.B.S. (eds.): Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026). Lecture Notes in Computer Science, Springer (2026)
18. Heras Rivera, J.E., Low, D.K., Xiong, X., Ruzevick, J.J., Child, D.D., wai Yim, W., Kurt, M., Ben Abacha, A.: Core-bt: A multimodal radiology-pathology-text benchmark for robust brain tumor typing. vol. abs/2603.03618 (2026)
19. Hicks, S.A., Storås, A., Halvorsen, P., de Lange, T., Riegler, M.A., Thambawita, V.: Overview of imageclefmedical 2023 – medical visual question answering for gastrointestinal tract. In: CLEF2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 2023)
20. Hicks, S.A., Storås, A., Halvorsen, P., Riegler, M.A., Thambawita, V.: Overview of ImageCLEFmedical 2024 – medical visual question answering for gastrointestinal tract. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 2024)
21. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain,

J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), vol. 11438. LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
22. Ivanovici, M., Baicoianu, A., Plajer, I.C., Debu, M., Ștefan, F.M., Florea, C., Cațaron, A., Coliban, R.M., Popa, S., Oprisescu, S., Racoviteanu, A., Olteanu, G., Marandskiy, K., Ghinea, A., Kazak, A., Majercsik, L., Manea, A., Dogaru, L.: Ai4agri sentinel-2 brasov area 2020-2024 multi- spectral dataset for crop monitoring and identification (Dec 2024). https://doi.org/10.5281/zenodo.14283243, https://doi.org/10.5281/zenodo.14283243
23. Jha, D., Ali, S., Emanuelsen, K., Hicks, S.A., Thambawita, V., Garcia-Ceja, E., Riegler, M.A., de Lange, T., Schmidt, P.T., Johansen, H.D., Johansen, D., Halvorsen, P.: Kvasir-Instrument: Diagnostic and Therapeutic Tool Segmentation Dataset in Gastrointestinal Endoscopy. In: Proceedings of the International Conference on MultiMedia Modeling (MMM). pp. 218–229 (2021), https://doi.org/10.1007/978-3-030-67835-7_19
24. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., de Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: Proceeding of the International Conference on Multimedia Modeling (MMM). vol. 11962, pp. 451–462 (2020), https://doi.org/10.1007/978-3-030-37734-2_37
25. Kondepudi, A., Rao, A., Zhao, C., Lyu, Y., Harake, E.S., Banerjee, S., Joshi, R.S., Meißner, A., Hou, R., Jiang, C., Chowdury, A., Srinivasan, A., Athey, B., Gulani, V., Pandey, A., Lee, H., Hollon, T.C.: Health system learning achieves generalist neuroimaging models. CoRR **abs/2511.18640** (2025). https://doi.org/10.48550/ARXIV.2511.18640, https://doi.org/10.48550/arXiv.2511.18640
26. Pereira, J.A., Macêdo, D., Zanchettin, C., de Oliveira, A.L.I., do Nascimento Fidalgo, R.: PictoBERT: Transformers for next pictogram prediction. Expert Systems with Applications **202**, 117231 (2022)
27. Rückert, J., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Schmidt, C.S., Koitka, S., Pelka, O., Ben Abacha, A., García Seco de Herrera, A., Müller, H., Horn, P.A., Nensa, F., Friedrich, C.M.: ROCOv2: Radiology Objects in COntext version 2, an updated multimodal image dataset. Scientific Data **11**(1), 688 (2024). https://doi.org/10.1038/s41597-024-03496-6
28. Stanciu, D., Radu, A., Bolborici, R., Negru, M., Ene, A., Vasilescu, V., Nicolae, A.A., Ionescu, B., Ştefan, L., Constantin, M., Dogariu, M., Andrei, A.: Overview of imageclef 2026 deepfake task: Multimodal detection and generation of deepfakes. In: CLEF 2026 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany (September 21–24 2026)
29. Sun, Z., Jagtiani, M., Yim, W., Xia, F., Gunn, M., Yetisgen, M., Ben Abacha, A.: Radar: A multimodal benchmark for 3d image-based radiology report review. CoRR **abs/2603.06681** (2026), https://arxiv.org/abs/2603.06681
30. Team, Q.: Qwen3 technical report. https://arxiv.org/abs/2505.09388 (2025)
31. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., Xu, Y., Wei, M., Wang, W., Ma, S., Wei, F., Yang, J., Li, C., Gao, J., Rosemon, J., Bower, T., Lee, S., Weerasinghe, R., Wright, B.J., Robicsek, A., Piening, B., Bifulco, C., Wang, S., Poon, H.: A whole-slide foundation model for digital pathology from real-world data. Nature (2024)