Overview of ImageCLEF 2025: Multimedia Retrieval in Medical, Social Media and Content Recommendation Applications

Bogdan Ionescu[1], Henning Müller[2], Dan-Cristian Stanciu[1], Alexandra-Georgiana Andrei[1], Ahmedkhan Radzhabov[3], Yuri Prokopchuk[4], Liviu-Daniel Ştefan[1], Mihai Gabriel Constantin[1], Mihai Dogariu[1], Vassili Kovalev[3,4], Hendrik Damm[5], Johannes Rückert[5], Asma Ben Abacha[6], Alba G. Seco de Herrera[7], Christoph M. Friedrich[5], Louise Bloch[5], Raphael Brüngel[5], Ahmad Idrissi-Yaghir[5], Henning Schäfer[8], Cynthia Sabrina Schmidt[8], Tabea M. G. Pakull[8], Benjamin Bracke[5], Obioma Pelka[5], Bahadır Eryılmaz[9], Helmut Becker[9], Wen-Wai Yim[6], Noel Codella[6], Roberto Andres Novoa[10], Josep Malvehy[11], Dimitar Dimitrov[12], Rocktim Jyoti Das[13], Zhuohan Xie[14], Ming Shan Hee[15], Preslav Nakov[14], Ivan Koychev[12], Steven A. Hicks[15], Sushant Gautam[15], Michael A. Riegler[15], Vajira Thambawita[15], Pål Halvorsen[15], Diandra Fabre[17], Cécile Macaire[17], Benjamin Lecouteux[17], Didier Schwab[17], Martin Potthast[18], Maximilian Heinrich[19], Johannes Kiesel[20], Moritz Wolter[17], Sharat Anand[19], and Benno Stein[19]

[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] Belarus State University, Belarus
[4] Belarusian National Academy of Sciences, Belarus
[5] University of Applied Sciences and Arts Dortmund, Germany
[6] Microsoft, USA
[7] University of Distance Education (UNED), Spain
[8] Institute for Transfusion Medicine, University Hospital Essen, Germany
[9] Institute for Artificial Intelligence in Medicine, University Hospital Essen
[10] Stanford University, USA
[11] Hospital Clinic of Barcelona, Spain
[12] Sofia University "St. Kliment Ohridski", Bulgaria
[13] Indian Institute of Technology, Delhi
[14] Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
[15] SimulaMet, Norway
[16] Université Grenoble Alpes, LIG, France
[17] University of Kassel, hessian.AI, and ScaDS.AI, Germany
[18] Bauhaus-Universität Weimar, Germany
[19] GESIS – Leibniz Institute for the Social Sciences, Germany
[20] Leipzig University, Germany

Abstract. This paper presents an overview of the ImageCLEF 2025 lab, which was organized within the Conference and Labs of the Evaluation Forum – CLEF Labs 2025. ImageCLEF is an ongoing evaluation event that started in 2003, promoting the evaluation of technologies for

annotation, indexing, and retrieval of multimodal data and aiming to provide access to large collections of data across a veriety of scenarios, domains and contexts. In 2025, the 23rd edition of ImageCLEF consists of four main tasks: (i) the *Medical* task, comprised of four sub-tasks, approaching a wide array of problems in the medical field, like concept detection, caption prediction, explainability assessment in radiology images, evaluating the veracity of GAN-generated 3D CT scans, providing a segmentation and answers to close-ended questions regarding dermatology images, or visual question answering and synthetic image generation involving gastrointestinal images, (ii) a new *Multimodal Reasonin*g task, involving answering multiple-choice questions in 13 different languages, covering a wide range of subjects and difficulty levels , (iii) the *ToPicto* task, which focuses on converting either text or speech into a meaningful sequence of pictograms and (iv) the *Argument-Image* task, which explores the augmentation of arguments using images, by either retrieval or synthetic generation. This edition of the ImageCLEF benchmark attracted 193 teams that registered to the different tasks, of which 56 finished the challenges. This resulted in 493 submitted runs and a total of 45 working note papers. Overall, this year's edition has been very successful, with the biggest number of teams, submissions and working notes papers since 2019.

Keywords: Medical image processing · Medical image caption analysis · Medical concept prediction · Visual question answering · Generative Adversarial Networks · Synthetic Data Generation · Image Segmentation · Pictogram communication · Multilingual · Image Retrieval · ImageCLEF

1 Introduction

Since its inception in 2003 [7], ImageCLEF[21] has evolved to be one of the most prominent evaluation initiatives, promoting the evaluation of technologies for indexing, retrieval of visual data, and facilitating access to large image collections across a large variety of domains. This paper presents an overview of the 2025 edition of the lab, part of the Conference and Evaluation Forum- CLEF Labs 2025 [5].

Starting from just 4 participants in 2003, the impact of this evaluation campaign grew over the years, amassing thousands of participants, and runs over the years. The ever-growing impact of ImageCLEF -and also CLEF more broadly - has been assessed in [24, 25], and is further evidenced by the number of results for the term "ImageCLEF" returned by Google Scholar, with over 7900 mentions[22] to date. Over the years, ImageCLEF has always adopted emerging trends, adding tasks of interest, with some of the more recent trends included in this year's tasks being the evaluation of Vision-Language Models, Synthetic Image Generation or AI model Explainability.

[21] `http://www.imageclef.org/`

[22] https://scholar.google.com/scholar?q=ImageCLEF

This edition of ImageCLEF features 4 main tasks, with a large diversity of sub-tasks: ImageCLEFMedical, MultimodalReasoning, ImageCLEFtoPicto and Image Retrieval/Generation for Arguments.

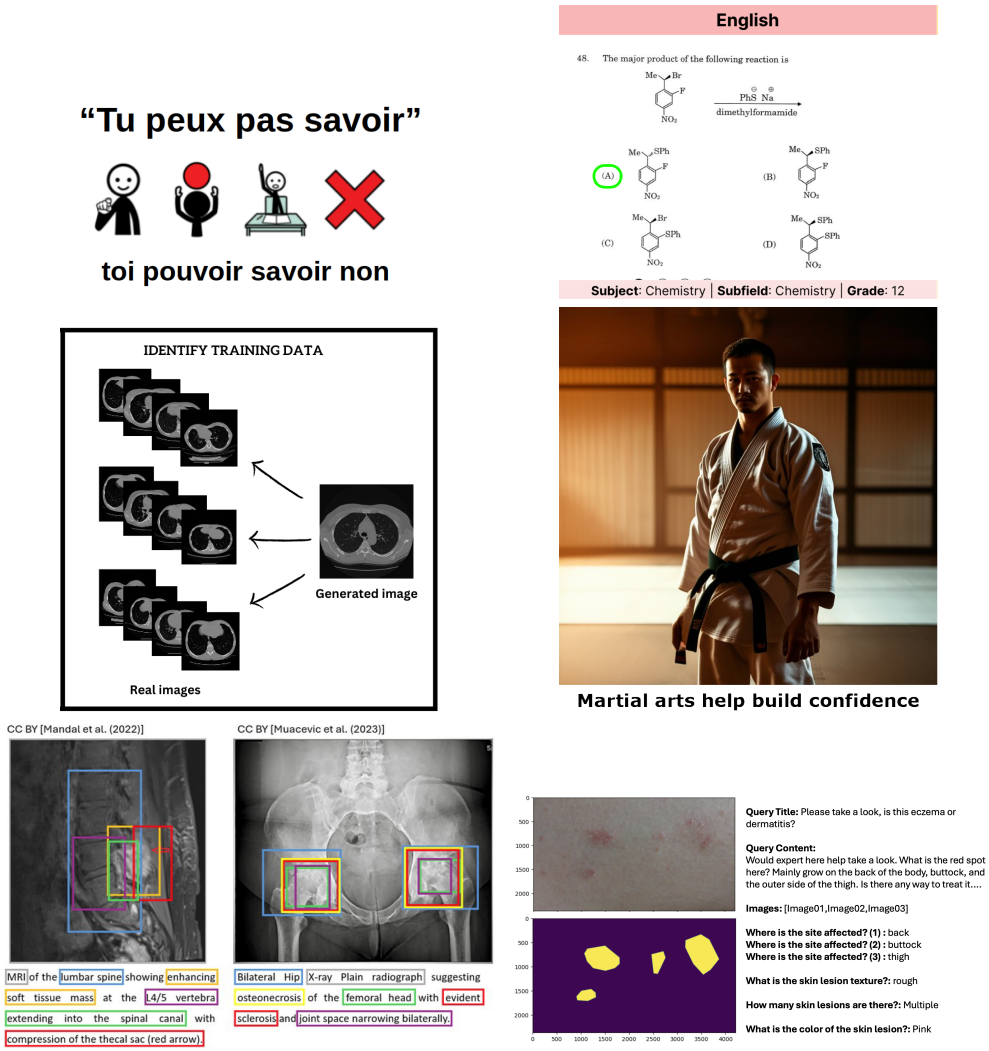

Fig. 1: Sample images from (left to right, top to bottom): ToPicto, Multimodal-Reasoning , the GANs task, Image Retrieval/Generation for Arguments, the Caption task and MEDIQA-MAGIC

2 Overview of Tasks and Participation

ImageCLEF 2025 consists of four main tasks to cover a *diverse range* of *multimedia retrieval*, *medical* applications. It followed the 2019 tradition [14] of diversifying the use cases [21, 23, 29, 22, 13, 3]. The 2025 tasks are presented as follows:

- **ImageCLEFmedical**. Since 2004, the ImageCLEF benchmarking initiative has included medical tasks. By 2018, however, although nearly all tasks were medical, there was limited interaction between them. Therefore, starting in 2019, the medical tasks were consolidated into a single medical task, with multiple subtasks, each handling a separate problem. This approach fostered synergies between the different domains. The medical task features four subtasks, described below:
  - *Caption*: The 2025 edition marks the nineth iteration of the medical captioning task [8]. This year, the task was expanded to three subtasks: the returning concept detection and caption prediction subtasks, and a newly promoted official explainability subtask. The caption prediction subtask focuses on composing coherent captions for radiology images, concept detection on identifying relevant UMLS concepts within those images, while the new explainability subtask requires participants to provide human-interpretable justifications for their model's predictions.
  - *GANs*: This is the third edition of the task [2, 3, 1]. The objective is to investigate whether synthetic medical images generated by deep generative models, such as GANs and Diffusion models, retain identifiable traces of the real data used during training. Addressing critical privacy and security concerns, the task includes two subtasks: detecting whether specific real images were used in GAN training, and attributing synthetic images to the correct training subset. The goal is to assess the potential for data leakage through "fingerprints" embedded in synthetic outputs
  - *MEDIQA-MAGIC*: The second edition for the MEDIQA-MAGIC [26] task builds on last year's challenges [28] using an expanded multimodal dermatology dataset. Participants receive clinical narratives with related images and must complete two subtasks: (1) segmenting areas showing dermatological issues, and (2) answering closed-ended clinical questions based on the provided context. Test sets are annotated by at least three annotators. Questions and options are available in both English and Chinese.
  - *MEDVQA*: Analysis of gastrointestinal (GI) images and videos continues to be an active research area in both the medical and computer science communities. Traditionally, most methods have focused on images as a single modality. The MEDVQA challenge [12] extends this by introducing a multimodal task that combines image and text data for visual question answering (VQA), targeting the field of GI endoscopy. This year, the challenge includes two subtasks. The first subtask focuses on answering clinical questions associated with specific images. The second subtask

involves generating synthetic GI images based on prompts describing anatomical landmarks, visual features, and other relevant findings.
- **MultimodalReasoning**. This is the first edition of the task. The objective is to assess how effectively vision-language models can reason over complex visual and textual exam content. Participants were provided with an image of a question, including answer options and metadata describing the type of visual content in the image. Their task was to select a single correct answer from a set of three to five options. The task combines a vision-language question answering problem with multilingualism, with subtasks in 14 languages, as well as a multilingual challenge.
- **ToPicto**. This second edition of the ToPicto task challenges participants to automatically generate pictogram translations from written text and speech. Participants trained models on a novel multimodal dataset comprising three aligned corpora. These resources include parallel speech, text, and pictogram sequences across various domains (medical, general) and settings (read or spontaneous speech) to support robust cross-modal translation.
- **Touché-Argument-Images**. This is the fourth edition of the task, which challenges participants to convey an argument through a single image. Given a central claim—such as "Martial arts help build confidence" —participants must either retrieve a relevant image from a dataset or generate one using an image synthesis tool. The task was conducted in collaboration with the Touché Lab. Further details are available in their overview paper [15].

Table 1: Key figures regarding participation in ImageCLEF 2025.

Task	Groups that submitted results	Submitted runs	Submitted working notes
GANs	15	105	9
Caption	11	149	10
MultimodalReasoning	11	129	11
MEDIQA-MAGIC	8	82	7
MEDVQA	6	17	5
ToPicto	3	7	2
Argument-Images	2	4	1

In order to participate in the evaluation campaign, the research groups had to register by following the instructions on the ImageCLEF 2025 web page[23]. Since 2024, we used our own registration and submission platform[24]. The Ai4Media platform allows for registration, data download, and automatic submission and evaluation of runs. Similar to previous editions, the participants were required to submit and sign the End User Agreement to access the datasets and submit runs.

[23] `https://www.imageclef.org/2025/`
[24] `https://ai4media-bench.aimultimedialab.ro/`

Following a drop in participation in 2016, interest in ImageCLEF rebounded in 2017 and 2018, with another increase observed in 2019. In 2018, 31 teams completed the tasks, resulting in 28 working notes papers. Participation peaked in 2019 with 63 teams and 50 submitted papers. In 2020 and 2021, 40 and 42 teams, respectively, completed the tasks, with 30 working notes received in 2021. In 2022, a decline followed, with 28 participating teams and 26 articles. However, 2023 marked a revival, with 47 teams submitting results and 39 working notes received. The 2024 edition attracted 26 teams, with a total of 34 working notes received. This year has seen the biggest number of participants since 2019, with with 55 teams submitting results and 45 working notes submitted. Table 1 presents the overall participation statistics for this year's competition.

The following sections present the tasks, outlining the most important points, like general objectives, description, data sets and results, in short overviews. A detailed review of the received submissions for each task is provided with the task extended overview in the CLEF 2025 working notes: Caption [8], GANs [1], MEDIQA-MAGIC [26], MEDVQA [12], MultimodalReasoning [11], ToPicto [19].

3 The Caption Task

The 9th edition of the ImageCLEFmedical Caption task continues to benchmark automatic systems for radiology image understanding. Building on previous years, the 2025 edition introduced two major changes: the promotion of explainability to a fully graded subtask alongside concept detection and caption prediction, and a revised, holistic evaluation methodology for the generated captions. The task attracted 11 participating teams who submitted a total of 149 graded runs.

3.1 Task Setup

Participants were invited to take part in up to three subtasks:

- **Concept Detection:** Systems were required to predict a set of Unified Medical Language System® (UMLS) concepts present in an image. Performance was measured by the F1-score.
- **Caption Prediction:** Systems had to generate a coherent, full-sentence caption for an image. A key update for 2025 was the evaluation via a composite score, averaging six metrics (including BERTScore, ROUGE-1, and AlignScore) to assess both relevance and factuality.
- **Explainability:** For a small, pre-selected subset of images, teams had to provide a human-interpretable explanation (e.g., a heat-map or bounding boxes) justifying their model's caption. Submissions were manually rated by a radiologist on a 1-5 Likert scale for criteria such as coherence and clinical relevance.

3.2 Data Set

The task used an enlarged and updated version of the ROCOv2 dataset, containing images and captions from the PubMed Central® Open-Access subset. The final collection comprised 116,635 images, split into training (80,091), validation (17,277), and test (19,267) sets. A key novelty for 2025 was the introduction of the optical coherence tomography (OCT) imaging modality, which was retrospectively annotated for the entire corpus. UMLS concepts were extracted using MedCAT and subsequently filtered to ensure a high-quality label space. For the explainability subtask, a dedicated set of 16 images was manually selected by a radiologist to cover all modalities.

3.3 Participating Groups and Submitted Runs

The 2025 task attracted 80 registered research groups, from which 11 internationally diverse teams ultimately submitted 149 graded runs. Ten teams submitted working notes (3 recurring teams). Participation was highest in the caption prediction subtask (8 teams, 98 runs), followed by concept detection (9 teams, 51 runs), and the new explainability subtask (2 teams, 2 runs). Six of the teams competed in both of the main subtasks.

3.4 Results

In the concept detection task, top-performing teams continued to rely on ensembles of Convolutional Neural Networks (CNNs). The winning submission from the AUEB NLP Group [6] exemplifies this mature approach. As shown in Table 2, while the leading F1-scores were competitive, there was a general decrease in primary scores compared to previous years. This is attributed to the increased difficulty and diversity of the 2025 dataset, particularly with the introduction of the OCT modality.

Table 2: Best-run performance of participating teams in the concept detection subtask, ranked by primary F1-score.

Group Name	F1 Secondary	F1 Rank (secondary)
AUEB NLP Group	**0.5888**	**0.9484**	1 (1)
DeepLens	0.5766	0.9299	2 (2)
mapan	0.5660	0.9298	3 (3)
UIT-Oggy	0.5613	0.9104	4 (4)
DS4DH	0.5225	0.8672	5 (6)
sakthiii	0.4003	0.9082	6 (5)
JJ-VMed	0.3982	0.8329	7 (7)
UMUTeam	0.2398	0.5377	8 (8)
LekshmiscopeVIT	0.1494	0.2298	9 (9)

The caption prediction subtask saw a clear and universal shift towards fine-tuning large Vision-Language Models (VLMs). The winning team, UMUTeam [20],

used a fine-tuned BLIP architecture to achieve the best overall score. The results from the new composite metric, summarized in Table 3, highlight that while top systems could generate highly relevant captions, achieving high factuality scores proved challenging for all participants. Notably, a baseline model using an off-the-shelf instruction-tuned LLM (Llama 4 Scout) performed competitively, placing in the middle of the rankings and even outperforming some submissions on specific metrics.

Table 3: Best-run performance of participating teams in the caption prediction subtask, ranked by the new composite Overall score.

Group Name	Overall	Relevance	Factuality	Rank (Rel./Fact.)
UMUTeam	**0.3432**	**0.5268**	**0.1596**	1 (1/1)
DS4DH	0.3362	0.5174	0.1549	2 (2/2)
AI Stat Lab	0.3229	0.5089	0.1369	3 (3/3)
UIT-Oggy	0.3211	0.5076	0.1346	4 (4/4)
AUEB NLP Group	0.3068	0.4759	0.1377	5 (6/5)
JJ-VMed	0.3043	0.4922	0.1165	6 (5/6)
sakthiii	0.2746	0.4481	0.1011	7 (7/7)
CS_Morgan	0.2315	0.3717	0.0917	8 (8/8)
Baseline (Llama 4 Scout)	0.3101	0.5073	0.1128	

Finally, the inaugural explainability task saw participation from two teams. The manual evaluation by a radiologist revealed that both teams generated plausible visualizations (e.g., bounding boxes). However, these explanations were created using external, post-hoc models and did not provide insight into the internal reasoning of the captioning models themselves.

3.5 Lessons Learned and Next Steps

The 2025 task yielded several important insights for the field. First, while complex pipelines like Retrieval-Augmented Generation (RAG) were explored, the top results in captioning came from direct and robust fine-tuning of strong VLM backbones, suggesting that pipeline complexity can be a weakness if components are not perfectly optimized. Second, the new composite score proved to be a successful evolution, giving a more balanced view of caption quality. Its results clearly pinpoint clinical factuality as the next major frontier for research, as relevance scores now consistently outperform factuality scores.

The most critical lesson came from the new explainability task. The reliance of participants on post-hoc, external models highlights a need for the community to focus on developing and evaluating model-intrinsic explanation methods (e.g., attention maps, GradCAM) that can surface the actual features the generative model used, which is essential for building clinical trust.

Based on these lessons, the following steps are planned for the 2026 challenge:

– **Mature the Explainability Task:** Future guidelines will strongly encourage the submission of model-intrinsic explanations to better align the task with the goal of building trustworthy AI.
– **Expand and Enrich the Dataset:** The dataset will be expanded again with new articles. To address multilinguality, previously omitted non-English captions will be machine-translated and incorporated. Furthermore, for images that lack a caption, a baseline will be generated using the wider article context, providing a new type of training data.

4 The GANs Task

4.1 Task Setup

The third edition of the ImageCLEFmedical 2025 GANs task [1] builds on the first two editions [2, 3], continuing the exploration of privacy and security concerns in synthetic medical imaging. The task is organized into two subtasks:

– **Subtask 1: Detect Training Data Usage** – Participants had to determine whether specific real images were used in training a Generative Adversarial Network (GAN) that produced a given set of synthetic images.
– **Subtask 2: Identify Training Data Subsets** – The goal was to link each synthetic image generated by a Diffusion model to its corresponding training subset among five predefined groups.

4.2 Data Set

The benchmarking datasets used in the GANs Task focused on computed tomography (CT) images spanning several anatomical regions relevant to medical imaging research. In Subtask 1, the data comprised thoracic CT axial slices originating from patients with lung tuberculosis. These slices presented a wide range of visual characteristics, from normal appearing lungs to scans with severe pulmonary lesions. Synthetic images were generated using a GAN trained on a subset of these real thoracic scans. In Subtask 2, the dataset extended to include cervical and abdominal CT slices, in addition to thoracic regions. This subtask leveraged a Diffusion-based generative model. All images, both real and synthetic, were standardized in format and resolution, enabling consistent evaluation across both subtasks while reflecting realistic clinical heterogeneity in organ appearance and pathology.

4.3 Participating Groups and Submitted Runs

Overall, 41 teams registered for our task. Of these, 14 teams completed the first subtask by submitting runs, and 4 teams completed the second subtask. In total, 9 teams submitted working notes papers. Notably, 2 teams participated in both subtasks, including the task organizing team. The continued interest in

the first subtask, now in its third edition, highlights its ability to attract more participating teams.

Each participating team was allowed to submit up to 10 runs per task. We received a total of 105 submitted runs: 91 for Subtask 1 and 14 for Subtask 2.

4.4 Results

The rankings for Subtask 1 are shown in Table 4, and those for Subtask 2 are presented in Table 5, limited to the teams that described their methods by submitting working notes papers.

Table 4: Results of participant submissions and their results for Subtask 1: Detect Training Data Usage.

#	Participant	Run ID	Cohen's kappa	Accuracy	Precision	Recall	F1
1	Neural Nexus	1878	**0.148**	0.574	0.5698	0.604	0.5864
2	zhouyijiang1	1803	**0.136**	0.568	0.5582	0.652	0.6015
3	zhouyijiang1	1804	**0.136**	0.568	0.5582	0.652	0.6015
4	zhouyijiang1	1873	**0.136**	0.568	0.5582	0.652	0.6015
5	zhouyijiang1	1802	**0.132**	0.566	0.5537	0.68	0.6104
6	zhouyijiang1	1801	**0.128**	0.564	0.55	0.704	0.6175
7	Neural Nexus	1880	**0.072**	0.536	0.5542	0.368	0.4423
8	taotaozi	1359	**0.064**	0.532	0.5597	0.3	0.3906
9	taotaozi	1367	**0.044**	0.522	0.5505	0.24	0.3343
10	AIMultimediaLab*	1696	**0.036**	0.518	0.5162	0.572	0.5427
11	taotaozi	1364	**0.032**	0.516	0.6	0.096	0.1655
12	Neural Nexus	1881	**0.032**	0.516	0.5222	0.376	0.4372
13	taotaozi	1360	**0.032**	0.516	0.5128	0.64	0.5694
14	Neural Nexus	1877	**0.028**	0.514	0.5164	0.44	0.4752
15	taotaozi	1366	**0.02**	0.51	0.5069	0.732	0.599
16	Neural Nexus	1872	**0.016**	0.508	0.5182	0.228	0.3167
17	Medhastra	1288	**0.016**	0.508	0.5078	0.52	0.5138
18	taotaozi	1368	**0.012**	0.506	0.5092	0.332	0.4019
19	Challengers	1811	**0.012**	0.506	0.5062	0.492	0.499
20	ZOQ	1427	**-0.016**	0.492	0.4905	0.412	0.4478
21	Neural Nexus	1879	**-0.024**	0.488	0.4732	0.212	0.2928
22	Neural Nexus	1882	**-0.028**	0.486	0.4646	0.184	0.2636
23	ZOQ	1355	**-0.032**	0.484	0.4904	0.82	0.6138
24	SCOPE VIT Visioneers	1160	**-0.032**	0.484	0.4831	0.456	0.4691
25	Challengers	1779	**-0.032**	0.484	0.4355	0.108	0.1731
26	AIMultimediaLab*	1492	**-0.044**	0.478	0.4829	0.62	0.5429
27	ZOQ	1330	**-0.068**	0.466	0.4822	0.92	0.6327
28	ZOQ	1794	**-0.068**	0.466	0.4822	0.92	0.6327
29	taotaozi	1369	**-0.096**	0.452	0.4657	0.652	0.5433
30	Challengers	1778	**-0.116**	0.442	0.4461	0.48	0.4624
31	ZOQ	1356	**-0.132**	0.434	0.3862	0.224	0.2835
32	Challengers	1776	**-0.176**	0.412	0.3764	0.268	0.3131
33	Challengers	1777	**-0.176**	0.412	0.3764	0.268	0.3131

Subtask 1, which asked participants to detect whether a specific real image was used in the training process of a GAN, proved to be highly challenging. Despite the use of a wide range of techniques, including Siamese neural networks, contrastive learning, supervised classifiers, clustering approaches, and advanced Vision Transformer-based architectures, overall performance remained modest.

The best result was achieved by the Neural Nexus team, whose ViT-based autoencoder pipeline reached a Cohen's kappa score of 0.148, with others, such as zhouyijiang1, achieving slightly lower but consistent scores around 0.13. Most submissions hovered close to or below zero, suggesting that models struggled to extract any reliable signal beyond chance. This low inter-rater agreement, measured via Cohen's kappa, underscores the difficulty of the task and possibly reflects that the GAN used in this edition was effective in generating images without obvious "fingerprints" of the training data.

In contrast, Subtask 2, which required participants to attribute synthetic images to their corresponding training subset of real data, showed substantially higher performance. Multiple teams reached classification accuracies above 98%, demonstrating that although image-level membership inference remains difficult, coarse-grained attribution at the dataset level is more tractable. The highest performance was recorded by SDVAHCS/UCSD, whose ensemble-based approach using EfficientNet architectures and pseudo-labeling,achieved up to 98.8% accuracy. Medhastra, using a simpler ResNet-18 classifier, also performed well with an accuracy of 94.84%, proving that even relatively lightweight models can be effective for this task when well-tuned.

Table 5: Results of participant submissions and their results for Subtask 2: Identify Training Data Subsets.

#	Participant	Run ID	Accuracy	Precision	Recall	F1	Specificity
1	AIMultimediaLab*	1396	**0.9904**	0.9904	0.9904	0.9904	0.9972
2	SDVAHCS/UCSD	1782	**0.988**	0.9882	0.988	0.9881	0.9969
3	SDVAHCS/UCSD	1871	**0.988**	0.9882	0.988	0.9881	0.9969
4	SDVAHCS/UCSD	1883	**0.988**	0.9882	0.988	0.9881	0.9969
5	SDVAHCS/UCSD	1426	**0.9878**	0.9881	0.9878	0.988	0.9969
6	SDVAHCS/UCSD	1425	**0.9708**	0.9716	0.9708	0.9711	0.9931
7	Medhastra	1287	**0.9484**	0.9504	0.9484	0.9487	0.9879
8	AIMultimediaLab*	1268	**0.5236**	0.5982	0.5236	0.5327	0.8799
9	AIMultimediaLab*	1269	**0.4913**	0.5822	0.4913	0.4934	0.8744
10	AIMultimediaLab*	1271	**0.4904**	0.5691	0.4904	0.4832	0.8753
11	AIMultimediaLab*	1267	**0.4112**	0.4645	0.4112	0.3945	0.8547

4.5 Lessons Learned and Next Steps

The 2025 ImageCLEFmedical GANs Task provided valuable insights into the privacy implications of synthetic medical image generation. Subtask 1, which focused on detecting whether specific real images were used to train a GAN, proved especially challenging. Despite using advanced techniques, participant systems performed close to random, with the best Cohen's kappa reaching only 0.148. This suggests that, under the conditions of this task, the proposed GAN generated images that were effectively free from directly traceable "fingerprints". While this is promising from a privacy perspective, it also reflects the limitations of current detection methods.

Subtask 2, in contrast, yielded significantly better results. Several teams achieved accuracies above 98%, showing that dataset-level attribution remains feasible. Successful methods employed supervised classification, deep feature embeddings, and semi-supervised learning strategies like pseudo-labeling. These results imply that while image-level membership inference is difficult, generative models may still retain broader signals from the source data distributions.

The differing outcomes of the two subtasks underscore the need to evaluate both model type and attribution level when assessing privacy risks. Moving forward, future task editions aim to explore new modalities, expand to multimodal generative data, and introduce more challenging attribution and privacy-preservation tasks.

5 The MEDIQA-MAGIC Task

In the second MEDIQA-MAGIC task [26], we extend on the previous year's dataset [31] and challenges [28, 27] based on multimodal dermatology response generation. In this edition, participants were asked to identify areas of interest in an image based on the patient's query, e.g. the rash on an arm, as well as provide answers to structured closed-ended questions, e.g. is there single or multiple lesions. These are critical subtasks that can be used to improve end-to-end free text response generation, the subject of the original 2024 challenge.

5.1 Task Setup

Similar to the previous edition, participants were given a clinical narrative context along with accompanying images. The task was divided into two relevant sub-parts: (i) segmentation of dermatological problem regions, and (ii) providing answers to closed-ended questions. The questions, answers, and answer options were given in both English and Chinese.

In the first sub-task, given each image and the clinical history, participants are tasked generating segmentations of the regions of interest for the described dermatological problem. The expected outputs are binary image files with the same size as the original image. To leverage multiple gold standard masks for segmentation, we use the majority vote by pixel as the gold standard for microscore calculations of Jaccard and Dice Index. However, we also calculate the mean of the per-instance max and mean.

In the second sub-task, participants were given a patient dermatological query, its accompanying images, as well as a closed-question with accompanying choices – the task is to select the correct answer to each closed question. Because the same dermatological problem may have multiple sites, there may be related questions (e.g. what is the size of the affected area for location 1, what is the size of the affected area for location 2). In these cases, the answers to the same related questions are collated together. Partial credit is given when there are partial matches to gold. The exact code can be found here: github.com/wyim/ImageCLEF-MAGIC-2025.

5.2 Data Set

The dataset was created by using real consumer health users' queries and images; the question schema was created in collaboration with two certified dermatologists. In total closed question schema - a comprehensive list of clinically relevant, patient-facing questions for dermatological assessments included a total of 137 questions. For the challenge, we tested for total of 27 questions, for which were most common and can use both text and images to answer. These corresponded to 9 overall questions when related questions are grouped (e.g. anatomic region for affected area 1, anatomic region for affected area 2). The answers were labeled by at least 3 annotators: 2 medical scribe annotators, 1 biomedical informatics graduate student. Questions and answers were translated into Chinese by a native Chinese speaker. Full details can be found in our dataset paper [30]

Congruent with the MEDIQA-M3G edition [27], there was a total of 300, 56, and 100 instances for train, valid, and test splits respectively. Each query had on average 3 images.

5.3 Participating Groups and Submitted Runs

Fifty three teams registered for the event. A total of 56 completed valid runs across 6 teams were submitted. Table 6 provides a list of partipating teams and affiliations. This year's primary participants came from academia from United States, Vietnam and India.

Table 6: Participating Teams in the MEDIQA-MAGIC 2025 Challenge

Team	Institution	Affiliation
DS@GT	United States	Georgia Institute of Technology
H3N1	Vietnam	University of Information Technology
Kasukabe Defense Group	India	KLE technological university
Anastasia	Vietnam	Universiy of Information Technology
IReL, IIT(BHU)	India	Indian Institute of Technology(BHU)
KLE1	India	KLE Technological University
Oggy	Vietnam	University of Information Technology

5.4 Results

Table 7 shows results for the segmentation tasks. Despite different calculations of jaccard and dice metrics, both given identical rankings. Table 8 shows results for the segmentation task.

In the segmentation subtask, all four teams took a fine-tuning approach with differences in the exact models employed (e.g. TransUNet, ViT-B, CLIP). The Anatasia team enriched the dataset by performing image transformation techniques (e.g. rotations, contrast etc) and were able to achieve top performances after including data with all transformations. The IReL, IIT(BHU) team was the only team that attempted to incorporate textual features. Their strategy used

Table 7: Performance of the participating teams in the MEDIQA 2025 Subtask 1 on segmentation generation for dermatological problems. Duplicate submission scores are removed.

team	jaccard	dice
Anastasia	0.6458	0.7848
Anastasia	0.6113	0.7587
IReL, IIT(BHU)	0.5881	0.7407
KLE1	0.5410	0.7021
H3N1	0.5145	0.6794
Anastasia	0.3205	0.4855
Anastasia	0.3129	0.4766
Kasukabe Defense Group	0.1866	0.3145

Table 8: Performance of the participating teams in the MEDIQA 2025 Subtask 2 on closed question answering. Duplicate submission scores are removed.

team	CQID010	CQID011	CQID012	CQID015	CQID020	CQID025	CQID034	CQID035	CQID036	ALL
H3N1	0.7	0.89	0.77	0.91	0.69	0.97	0.45	0.86	0.58	0.76
H3N1	0.64	0.89	0.76	0.87	0.71	0.96	0.47	0.85	0.6	0.75
H3N1	0.67	0.74	0.72	0.93	0.69	0.98	0.49	0.87	0.62	0.75
H3N1	0.64	0.88	0.76	0.85	0.73	0.9	0.46	0.86	0.54	0.74
DS@GT MEDIQA-MAGIC	0.53	0.87	0.66	0.81	0.56	0.89	0.6	0.81	0.65	0.71
DS@GT MEDIQA-MAGIC	0.51	0.84	0.7	0.85	0.56	0.87	0.55	0.81	0.67	0.71
DS@GT MEDIQA-MAGIC	0.47	0.86	0.69	0.85	0.56	0.84	0.51	0.82	0.64	0.69
DS@GT MEDIQA-MAGIC	0.44	0.84	0.69	0.78	0.55	0.86	0.48	0.79	0.65	0.68
DS@GT MEDIQA-MAGIC	0.49	0.82	0.63	0.74	0.56	0.79	0.51	0.75	0.59	0.65
KLE1	0.51	0.63	0.75	0.57	0.63	0.56	0.39	0.74	0.35	0.57
KLE1	0.47	0.62	0.7	0.58	0.62	0.56	0.36	0.76	0.3	0.55
Kasukabe Defense Group	0.44	0.66	0.75	0.28	0.66	0.44	0.52	0.77	0.3	0.54
Kasukabe Defense Group	0.4	0.61	0.73	0.29	0.65	0.44	0.52	0.76	0.33	0.53
Kasukabe Defense Group	0.49	0.49	0.67	0.32	0.48	0.41	0.01	0.76	0.55	0.46
DS@GT MEDIQA-MAGIC	0.31	0.38	0.53	0.31	0.31	0.42	0.01	0.72	0.37	0.37
Oggy	0.08	0.26	0.45	0.3	0.02	0.35	0.02	0.03	0.48	0.22
IReL, IIT(BHU)	0	0.44	0.48	0.17	0.44	0	0.02	0	0	0.17

CLIP to embed both text and visual features then afterwards fed the combined feature vector into a binary classification to predict the mask. The remaining teams fine-tuned previously trained skin lesion segementation models; the H3N1 team use the DermoSegDiff model, whereas the KLE1 team fine-tuned a Multi-Scale Feature Fusion Network model. Though these models were trained for skin lesions, likely more fine-tuning was required to completely adapt the model to this new dataset.

In the closed question-answering subtask, the top two performing teams H3N1 and DSGT employed multi-step architectures, including both fine-tuned models and LLM API's and ensembling methods. The former separated that task into four parts: (1) preprocessing, (2) information enrichment via image captioning, (3) fine-tuning and external API calls, (4) ensembling models from the previous step. The latter similarly had several layers (1) LLM fine-tuning with different models e.g. Qwen and LLAMA, (2) reasoning layer over output of (1) using Gemini, and (3) and agent layer that additionally has a RAG to reference the LanceDB dermatology corpus. In contrast, the other remaining groups had similar approaches, which utilized encoders for the images and text, then

after fusing both text and image features, the network would eventually be fed into a classification layer.

5.5 Lessons Learned and Next Steps

In the segmentation task, the most successful system were able to use data augmentation generated through image transformation techniques (e.g. color contrast changes). This is promising as other teams did experiment with skin lesion segmentation specific models however were not able to achieve as high results – suggesting more data would be required to adapt those models. The use of textual inputs was only tested by one group, suggesting that this is an area for future exploration.

In the closed QA task, we found the best systems included multiple models fine-tuned for the task as well as some ensembling and aggregation. The use of multi-modal large language models were critically more successful than the suite of fine-tuned multimodal approaches which relied on a shared embedding representation then trained to fine-tune on the classification task. This could be because the current dataset is relatively small thus the important of the large language models' access to external information became a determining factor.

This edition implemented both subtasks simultaneously, simplifying organization but resulting in many repeat submissions with changes to only one subtask. Future improvements could include platform support for concurrent phases. Most submissions came from academia; future efforts will focus on expanding industry participation.

6 The MedVQA Task

The third edition of the MedVQA challenge at ImageCLEF continued to emphasize the application of image-based machine learning in gastrointestinal (GI) screening. In this edition, the scope of the challenge was broadened to include two key subtasks from the previous two years: visual question answering (VQA) and text-to-image synthesis. Five teams participated who submitted a total of eight runs. An overview can be seen in Figure 9.

6.1 Task Setup

This year, we organized two subtasks to evaluate different capabilities of machine learning models in gastrointestinal (GI) imaging. Subtask 1 focused on

Table 9: An overview of the submissions to each task at MedVQA-GI.

	MedVQA 2023	MedVQA 2024	MedVQA 2025
# Registrations	26	22	31
# Task Participation	8	2	5
# Paper Submissions	6	2	5

answering clinical questions associated with annotated images from the challenge development dataset. Models were required to combine visual recognition, such as identifying tools, anatomical structures, or pathological features, with basic language understanding. Subtask 2 involved generating synthetic GI images from structured prompts describing anatomical locations, visual features, or the presence of tools. The goal was to produce images that resembled real endoscopic data and could be used for training or evaluating diagnostic models. Submissions for both subtasks were managed through a Hugging Face repository to ensure standardization and comparability across teams.

6.2 Data Set

The dataset used in this challenge is based on the publicly available HyperKvasir and Kvasir-VQA datasets, which include gastrointestinal endoscopy images from various anatomical sites and pathological conditions. For Subtask 1, the development set contained over 6,500 images from Kvasir-VQA, each annotated with one or more visual questions and corresponding answers. The questions fell into categories including Yes/No, Single-Choice, Multiple-Choice, Color, Location, and Count, targeting tasks like classification, reasoning, spatial localization, and attribute recognition. The test set introduced a distribution shift by using previously unreleased images from different sources. For Subtask 2, participants were given more than 2,000 image-caption pairs summarizing clinically relevant content such as anatomical features, abnormalities, or procedural elements. To increase variation and reduce overfitting, synthetic captions were also provided, generated using large language models and rule-based techniques. The test set for Subtask 2 was drawn from a separate, mixed-source dataset not included in the training data to support evaluation in unfamiliar clinical settings.

6.3 Results

Five teams submitted results for Subtask 1, and three teams participated in Subtask 2. In Subtask 1 (Table 10), IReL_IIT_BHU ranked first overall based on top scores across ROUGE and METEOR on the private set. UPS was the runner-up, with the highest BLEU score on the public set and strong performance on other metrics. In Subtask 2 (Table 11), CS_Morgan_Lab achieved the best overall performance, ranking first in FID, diversity, and FBD. IReL_IIT_BHU was the runner-up, with the highest agreement score and good scores on other metrics.

6.4 Lessons Learned and Next Steps

Most teams in Subtask 1 used transformer-based multimodal architectures, with Florence2 being the most common, fine-tuned using LoRA along with input augmentation and hyperparameter tuning. In Subtask 2, three teams submitted models based on fine-tuned variants of Stable Diffusion, also using LoRA. Although the generated images had high visual fidelity, prompt-image alignment

Table 10: Results for Task 1.

Team	Set	BLEU	R1	R2	RL	MET
UPS	Public	**0.24**	0.87	0.11	0.87	0.48
UPS	Private	0.22	0.88	0.11	0.88	0.49
IReL_IIT_BHU	Public	0.23	0.83	0.10	0.83	0.46
IReL_IIT_BHU	Private	0.22	**0.92**	**0.11**	**0.92**	**0.50**
MedPixel	Public	0.21	0.87	**0.12**	0.86	0.48
MedPixel	Private	0.18	0.91	0.11	0.90	**0.50**
CS_Morgan_Lab	Public	0.19	0.84	0.10	0.83	0.46
CS_Morgan_Lab	Private	0.18	0.90	0.10	0.90	0.49
Sagarmatha_Rangers	Public	0.15	0.81	0.10	0.80	0.44
Sagarmatha_Rangers	Private	0.16	0.88	0.10	0.88	0.49

Table 11: Results for Task 2.

Team	Set	Fid.	Agrmt.	Div.	FBD
CS_Morgan_Lab	Private	**0.0268**	0.7012	**0.7017**	**1539.31**
IReL_IIT_BHU	Private	0.2739	**0.7390**	0.6481	1694.97
MedPixel	Private	0.2725	0.7329	0.6722	1694.00

was inconsistent, and clinically accurate features were often missing. While more teams registered than last year, final submission numbers remained low. Subtask 2 may benefit from improved baselines, clearer instructions, and more human-in-the-loop evaluation.

7 The MultimodalReasoning task

7.1 Task Setup

The task focused on visual question answering for multiple-choice questions with exactly one correct answer, where the answer options could also include visual content. It was conducted in two phases: an exploration phase, during which participants familiarized themselves with the publicly available training and validation data [9], followed by a test phase. In the test phase, images of the questions from the test dataset were released along with metadata describing the visual components within the questions. Participants submitted results to 14 different leaderboards: one multilingual leaderboard and 13 individual leaderboards, one for each language. Participants were allowed to make multiple submissions during this phase, but no feedback was provided. Final rankings were determined based on each participant's last submission at the end of the test phase.

7.2 Data Set

The dataset used in this challenge is based on the publicly available Exams-V dataset [9], which includes 20,932 multiple-choice questions across 20 school disciplines, covering 11 languages from 7 language families. Each question had from three to five answer options, with a single correct answer. The test set

introduced new questions from more recent graduate exams. Table 12 shows general statistics on the test set, which contains a total of 3,565 new questions. Additionally, three new languages were introduced, Urdu, Kazakh, and Spanish, challenging participants to explore approaches for zero-shot question answering.

Table 12: New test data statistics. Here, #**visual Q.** refers to questions with multimodal context and #**text Q.** refers to text-only questions. **Urdu, Kazakh, and Spanish* are new languages, with no training/validation data from Exams-V.

Language	ISO	Family	Grade	#Subj.	#Questions	#visual Q.	#text Q.
English	en	Germanic	12	1	512	62	450
Chinese	zh	Sino-Tibetan	12	4	407	0	407
German	de	Germanic	12	6	258	68	190
Italian	it	Romance	12	5	203	58	145
Arabic	ar	Semitic	10-12	4	222	164	58
Polish	pl	Slavic	12	7	259	104	155
Hungarian	hu	Finno-Ugric	12	6	247	30	217
Bulgarian	bg	Slavic	12	6	200	66	134
Croatian	hr	Slavic	12	5	203	58	145
Serbian	sr	Slavic	12	5	203	58	145
Urdu*	ur	Indo-Aryan	9-10	5	269	0	269
Kazakh*	kk	Turkic	11	4	243	84	159
Spanish*	es	Romance	12	10	339	209	130

7.3 Participating Groups and Submitted Runs

In the first edition of the Multimodal Reasoning task, 51 participants registered, with 11 teams participating in the test set, resulting in a total of 129 graded submissions. All 11 teams submitted working notes. The most popular leaderboards were English, Multilingual, and Chinese, with 10, 9, and 7 teams participating, respectively. Some teams participated in multiple leaderboards, with two teams submitting to all 14 and another two teams submitting to 13. Teams came from 5 different countries: Bulgaria, China, India, Egypt, and Pakistan.

7.4 Results

Table 13 shows the results on all 14 leaderboards for the Multimodal reasoning task. Participants significantly outperformed the baseline, except one team that opted for the same model as the baseline. The task proved to be of moderate difficulty, with some teams achieving over 90% accuracy. Team **seifahmed** excelled across the board, securing first place in 11 out of the 13 leaderboards they competed in.

7.5 Lessons Learned and Next Steps

In the first edition of the Multimodal Reasoning task, we observed a lot of interest, with registration numbers being similar to other established tasks under

Table 13: Results for the ImageCLEF 2025 Multimodal Reasoning task on all 14 leaderboards. Baseline system submitted by the organizers. In the case of equal scores, participants are assigned the same rank and ordered alphabetically. †Participants submitted as different teams, but wrote a single working notes paper as co-authors.

Rank	Team	Acc	Rank	Team	Acc	Rank	Team	Acc
Multilingual			**English**			**Bulgarian**		
1	seifahmed	0.8140	1	stormhunter44†	0.8965	1	heavyhelium†	0.9050
2	ymgclef	0.5994	2	seifahmed	0.8652	1	stormhunter44†	0.9050
3	lekshmiscopevit	0.5770	3	ayeshaamjad	0.8125	2	ymgclef	0.7750
4	bingezzzleep	0.5619	4	heavyhelium†	0.8086	3	bingezzzleep	0.7500
5	plutohbj	0.5226	5	ymgclef	0.5938	3	seifahmed	0.7500
6	deng113abc	0.5195	6	deng113abc	0.5371	4	plutohbj	0.7300
7	mhl2001	0.4418	7	bingezzzleep	0.5312	5	baseline	0.2450
8	yaozihang	0.4376	8	plutohbj	0.4922	6	elenat	0.2350
9	baseline	0.2701	9	mhl2001	0.4629	**German**		
10	elenat	0.2188	10	yaozihang	0.4570	1	seifahmed	0.8915
Kazakh			11	elenat	0.2520	2	ymgclef	0.7403
1	seifahmed	0.8148	12	baseline	0.2480	3	bingezzzleep	0.6860
2	ymgclef	0.5350	**Chinese**			4	plutohbj	0.6783
3	bingezzzleep	0.4938	1	seifahmed	0.8305	5	yaozihang	0.4961
4	plutohbj	0.4444	2	ayeshaamjad	0.6560	6	mhl2001	0.4922
5	baseline	0.2738	3	plutohbj	0.5921	7	baseline	0.3101
Polish			4	bingezzzleep	0.5799	**Urdu**		
1	seifahmed	0.8224	5	mhl2001	0.5553	1	seifahmed	0.8067
2	ymgclef	0.7181	6	ymgclef	0.5283	2	ymgclef	0.3941
3	bingezzzleep	0.5792	7	yaozihang	0.4791	3	bingezzzleep	0.3569
4	plutohbj	0.5251	8	baseline	0.2678	3	yaozihang	0.3569
5	baseline	0.2934	**Arabic**			4	baseline	0.3011
Italian			1	seifahmed	0.6757	**Croatian**		
1	seifahmed	0.9212	2	ayeshaamjad	0.4775	1	seifahmed	0.9507
2	bingezzzleep	0.6059	3	mhl2001	0.4730	2	bingezzzleep	0.6207
2	plutohbj	0.6059	4	ymgclef	0.4324	3	ymgclef	0.5764
3	ymgclef	0.6010	5	plutohbj	0.3514	4	plutohbj	0.5616
4	baseline	0.2414	6	bingezzzleep	0.3243	5	baseline	0.2709
Spanish			7	baseline	0.2703	**Serbian**		
1	seifahmed	0.7198	**Hungarian**			1	seifahmed	0.7143
2	ymgclef	0.6696	1	ymgclef	0.6518	2	bingezzzleep	0.6059
3	bingezzzleep	0.6608	2	bingezzzleep	0.5425	3	ymgclef	0.5468
4	plutohbj	0.5723	3	plutohbj	0.4696	4	plutohbj	0.5320
5	baseline	0.3156	4	mhl2001	0.3563	5	baseline	0.2365
			5	baseline	0.2348			

the same lab. Participating teams opted to use a combination of proprietary and open-source large VLMs, including Qwen2.5-VL, Gemini, SmolVLM, and Deepseek. The majority of approaches employed zero-shot or few-shot techniques and leveraged metainformation about visual elements. There were some fine-tuning submissions, but these generally underperformed, primarily due to the use of smaller models constrained by limited resources. The most widely used models were Gemini-2.5 and Qwen2.5-VL, with the former consistently outperforming across all leaderboards, showing that the most recent advances in reasoning models can compete on graduate exams with complex visual elements.

8 The ToPicto task

The second edition of the ToPicto task focuses on the automatic generation of pictogram translations from two input modalities: written text and speech. This challenge introduces a novel multimodal dataset to support the training of machine learning models in a cross-modal translation setting.

Compared to the first edition, the dataset has been significantly extended to include a wider variety of acoustic domains (from read to spontaneous speech) and a broader set of thematic domains, including both medical and everyday-life contexts. Participants were tasked with building models that can generalize effectively and perform robustly across these diverse conditions.

8.1 Task Setup

The ToPicto 2025 task consists of two sub-tasks: Text-to-Picto and Speech-to-Picto. Participants were allowed to submit to one or both sub-tasks, with a maximum of 10 submissions in total.

- **Subtask 1: From Text to Pictogram Sequence** – The Text-to-Picto sub-task focuses on the automatic generation of a corresponding sequence of pictogram terms from a French text.
- **Subtask 2: From Speech to Pictogram sequence** – The Speech-to-Picto sub-task focuses on two modalities: speech and pictograms, and aims to directly map a speech input to pictogram concepts.

8.2 Data Set

The benchmarking data are curated from three aligned multimodal corpora: Propicto-commonvoice, Propicto-orféo, and Propicto-eval [16–18]. Propicto-commonvoice is based on the French portion of CommonVoice v15 [4], containing 967 hours of read speech from 17,911 speakers, with pictogram sequences generated using the method described in [16]. Propicto-orféo, built from the CEFC corpus [10], consists of 233 hours of spontaneous speech from various domains and interaction types, with corresponding pictogram translations. Propicto-eval is a controlled evaluation set with multi-speaker read speech derived from children's stories, everyday scenarios, and medical texts, intended to assess model performance across different content domains.

8.3 Participating Groups and Submitted Runs

In 2024, a total of 16 teams participated in the ToPicto challenge, and four teams completed the Text-to-Picto task and submitted their results. In 2025, a total of 41 teams participated in the ToPicto challenge and registered for both tasks. Only three teams completed the Text-to-Picto task (with 4 runs), and one team completed the Speech-to-Picto task (with 2 runs). Two teams merged their contributions for final submissions, resulting in two working notes provided.

8.4 Results

Table 14 shows the different submission results. Note that majahj and indira collaborated on both tasks and submitted a single working note.

Table 14: Performance of participating teams in the ToPicto 2025 task. Scores for sacreBLEU, METEOR, and PictoER are reported and ordered by the highest sacreBLEU score.

Sub-Task	Team Name	SacreBLEU↑	METEOR↑	PictoER (%)↓	Rank
Text-to-Picto	majahj	76,98	88,66	13,48	1
	sudharshan07	69,01	85,09	18,56	2
	indira	52,41	74,50	29,23	3
	indira	37,72	64,61	42,70	4
Speech-to-Picto	majahj	62,87	73,41	29,49	1
	majahj	54,71	65,90	40,02	2

8.5 Lessons Learned and Next Steps

In 2025, we observed a significant increase in registrations; however, only two teams submitted final working notes. Both teams conducted their work seriously and presented interesting results on the fine-tuning of Large Language Models for the pictogram translation task. One team focused on analyzing the impact of model size and number of training epochs, while the other designed a lightweight architecture tailored to the source language. For the next steps, emphasizing the place of this task in the NLP domain and providing participants with more documentation on pictograms should attract a wider range of profiles and more diverse solutions for solving both proposed tasks.

9 Conclusion

This paper presents the overview of the ImageCLEF 2025 benchmarking campaign. We introduce the four main tasks, introducing interesting challenges in medicine (caption analysis, medical data generation and assessment, medical image segmentation for question answering, visual question answering), multimodal and multilingual question answering, pictogram to text and audio translation and image retrieval/generation for arguments.

The majority of the approaches by the participants were deep learning-based. In the ImageCLEF Medical-Caption task, the top-performing teams used Convolutional Neural Networks for the Concept detection sub-task, as well as Vision-Language Models, for the Caption prediction problem. For the GANs task, the participating teams used a wide array of machine learning models, with a Vision Transformer architecture achieving the best results on the first subtask.

For the second subtask, multiple teams achieved a performance over 98% using deep learning approaches. For MEDIQA-MAGIC, the participants relied on Large Language Models. In the MedVQA task, the majority of teams used transformer-based architectures for the first subtask, while Stable Diffusion models were widely used in the second one. For the newly introduced Multimodal Reasoning task, most of the approaches used a combination of proprietary and open-source Vision Language Models, while employing zero-shot or few shot techniques. For ToPicto, the translation of text to pictograms was done using Large Language Models.

ImageCLEF continues to serve as a platform for innovation, learning, and advancement across a wide range of fields. Future editions will focus on enhancing existing tasks, expanding into new domains, attracting more participants, and fostering experimentation and learning in emerging areas. Addressing current limitations—such as clarifying task descriptions, allocating resources effectively, refining evaluation metrics, and exploring new assessment methods and collaboration opportunities—will also be a priority. Our goal remains to continually raise the quality of this benchmarking campaign and contribute meaningfully to progress in the field.

Acknowledgements

The work of Louise Bloch, Raphael Brüngel and Benjamin Bracke was partially funded by a PhD grant from the University of Applied Sciences and Arts Dortmund (FH Dortmund), Germany. The work of Ahmad Idrissi-Yaghir, Tabea M. G. Pakull, Hendrik Damm, Henning Schäfer, Bahadir Eryilmaz, and Helmut Becker 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 and Ivan Koychev is partially funded by the EU NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project SUMMIT, No BG-RRP-2.004-0008. The ToPicto task was 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, led by Didier Schwab and hosted by the Grenoble INP Foundation, with sponsorship from the Artelia Group. The chair also receives support from the French government, managed 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), that distributes them under Creative Commons License BY-NC-SA. The work of Bogdan Ionescu, Alexandra Andrei, Dan-Cristian Stanciu is supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-2-0320, within PNCDI IV. The work of Liviu-Daniel Stefan was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-0049, within PNCDI IV. The work of Mihai Gabriel Constantin was supported by a grant of the Ministry of Research, Innovation and Digitization, CCCDI - UEFISCDI, project number PN-IV-P6-6.3-SOL-2024-0060, within PNCDI IV. This work was partly supported by the project GRESEL-UNED PID2023-151280OB-C22 funded by MICIU/AEI/ AEI 501100011033.

References

1. 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)
2. 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)
3. 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)
4. Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 4218–4222. European Language Resources Association, Marseille, France (May 2020)
5. Carrillo de Albornoz, Gonzalo, J., Plaza, L., de Herrera, A.G.S., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., Ferro, N.: Experimental IR meets multilinguality, multimodality, and interaction. proceedings of the sixteenth international conference of the CLEF association (CLEF 2025). In: Proceedings of the 15th International Conference of the CLEF Association, CLEF 2024. Lecture Notes in Computer Science, Madrid, Spain (September 9–12 2025)
6. Chatzipapadopoulou, A., Pantelidis, I., Charalampakos, F., Samprovalaki, M., Moschovis, G., Kaliosis, P., Dalakleidi, K., Pavlopoulos, J., Androutsopoulos, I.: AUEB NLP group at ImageCLEFmedical Caption 2025. In: CLEF2025 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain (2025)
7. Clough, P., Sanderson, M.: The CLEF 2003 cross language image retrieval task. In: Proceedings of the Cross Language Evaluation Forum (CLEF 2003) (2004)
8. Damm, H., Pakull, T.M.G., Becker, H., Bracke, B., Eryilmaz, B., Bloch, L., Brüngel, R., Schmidt, C.S., Rückert, J., Pelka, O., Schäfer, H., Idrissi-Yaghir, A., Abacha, A.B., de Herrera, A.G.S., Müller, H., Friedrich, C.M.: Overview of ImageCLEFmedical 2025 – medical concept detection and interpretable caption generation. In: CLEF 2025 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain (September 9–12 2025)
9. 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)
10. Debaisieux, J.M., Benzitoun, C., Deulofeu, H.J.: Le projet ORFEO: Un corpus d'études pour le français contemporain. Corpus **15**, 91–114 (Jun 2016). https://doi.org/10.4000/corpus.2936, `https://hal.science/hal-01449600`

11. Dimitrov, D., Hee, M.S., Xie, Z., Jyoti Das, R., Ahsan, M., Ahmad, S., Paev, N., Koychev, I., Nakov, P.: Overview of imageclef 2025 – multimodal reasoning. In: CLEF 2025 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain (September 9–12 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. 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)
14. 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)
15. Kiesel, J., Çöltekin, Ç., Gohsen, M., Heineking, S., Heinrich, M., Fröbe, M., Hagen, T., Aliannejadi, M., Erjavec, T., Hagen, M., Kopp, M., Ljubešić, N., Meden, K., Mirzakhmedova, N., Morkevičius, V., Scells, H., Zelch, I., Potthast, M., Stein, B.: Overview of Touché 2025: Argumentation Systems. In: de Albornoz, J.C., Gonzalo, J., Plaza, L., García Seco de Herrera, A., Mothe, J., Piroi, F., Rosso, P., Spina, D., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 16th International Conference of the CLEF Association (CLEF 2025). Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York (Sep 2025)
16. Macaire, C., Dion, C., Arrigo, J., Lemaire, C., Esperança-Rodier, E., Lecouteux, B., Schwab, D.: A multimodal French corpus of aligned speech, text, and pictogram sequences for speech-to-pictogram machine translation. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 839–849. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.76/
17. Macaire, C., Dion, C., Schwab, D., Lecouteux, B., Esperança-Rodier, E.: Approches cascade et de bout-en-bout pour la traduction automatique de la parole en pictogrammes. In: Balaguer, M., Bendahman, N., Ho-dac, L.M., Mauclair, J., G Moreno, J., Pinquier, J. (eds.) Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position. pp. 22–35. ATALA and AFPC, Toulouse, France (7 2024)
18. Macaire, C., Dion, C., Schwab, D., Lecouteux, B., Esperança-Rodier, E.: Towards speech-to-pictograms translation. In: Interspeech 2024. pp. 857–861 (2024)
19. Macaire, C., Fabre, D., Lecouteux, B., Schwab, D.: Overview of the 2025 ImageCLEFtoPicto task – investigating the generation of pictogram sequences from text and speech. In: CLEF2025 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain (September 9-12 2025)

20. Pan, R., Bernal Beltrán, T., García Díaz, J.A., Valencia-García, R.: UMUTeam at ImageCLEF 2025: Fine-tuning a vision-language model for medical image captioning and SapBERT-based reranking for concept detection. In: CLEF2025 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain (September 9-12 2025)
21. Popescu, A., Deshayes-Chossart, J., Schindler, H., Ionescu, B.: Overview of the imageclef 2022 aware task. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy (September 5-8 2022)
22. Rückert, J., Ben Abacha, A., G. Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Bracke, B., Damm, H., Pakull, T.M.G., Schmidt, C.S., Müller, H., Friedrich, C.M.: Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 9-12 2024)
23. Ştefan, L.D., Constantin, M.G., Dogariu, M., Ionescu, B.: Overview of imagecleffusion 2023 task - testing ensembling methods in diverse scenarios. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
24. Tsikrika, T., García Seco de Herrera, A., Müller, H.: Assessing the scholarly impact of ImageCLEF. In: CLEF 2011. pp. 95–106. Springer Lecture Notes in Computer Science (LNCS) (sep 2011)
25. Tsikrika, T., Larsen, B., Müller, H., Endrullis, S., Rahm, E.: The scholarly impact of CLEF (2000–2009). In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. pp. 1–12. Springer (2013)
26. Yim, W., Ben Abacha, A., Codella, N., Novoa, R.A., Malvehy, J.: Overview of the MEDIQA-MAGIC task at ImageCLEF 2025: Multimodal and generative telemedicine in dermatology. In: CLEF 2025 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain (September 9-12 2025)
27. Yim, W.w., Ben Abacha, A., Fu, Y., Sun, Z., Xia, F., Yetisgen, M., Krallinger, M.: Overview of the MEDIQA-M3G 2024 shared task on multilingual multimodal medical answer generation. In: Naumann, T., Ben Abacha, A., Bethard, S., Roberts, K., Bitterman, D. (eds.) Proceedings of the 6th Clinical Natural Language Processing Workshop. pp. 581–589. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024)
28. Yim, W., Ben Abacha, A., Fu, Y., Sun, Z., Yetisgen, M., Xia, F.: Overview of the MEDIQA-MAGIC task at ImageCLEF 2024: Multimodal and generative telemedicine in dermatology. In: Conference and Labs of the Evaluation Forum (2024)
29. Yim, W., Ben Abacha, A., Snider, N., Adams, G., Yetisgen, M.: Overview of the MEDIQA-Sum task at imageclef 2023: Summarization and classification of doctor-patient conversations. In: CLEF 2023 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 18-21 2023)
30. Yim, W., Fu, Y., Ben Abacha, A., Yetisgen, M., Codella, N., Novoa, R.A., Malvehy, J.: Dermavqa-das: Dermatology assessment schema (das) and datasets for closed-ended question answering and segmentation in patient-generated dermatology images. CoRR (2025)
31. Yim, W., Fu, Y., Sun, Z., Ben Abacha, A., Yetisgen-Yildiz, M., Xia, F.: Dermavqa: A multilingual visual question answering dataset for dermatology. In: CLEF2024 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France (September 9-12 2024)