Overview of ImageCLEFmedical 2025 – Medical Concept Detection and Interpretable Caption Generation

Hendrik Damm[1,2,*,†], Tabea M. G. Pakull[3,1,†], Helmut Becker[4], Benjamin Bracke[1], Bahadır Eryılmaz[4], Louise Bloch[1,2,4], Raphael Brüngel[1,2,4], Cynthia S. Schmidt[4], Johannes Rückert[1], Obioma Pelka[4,5], Henning Schäfer[3,1,4], Ahmad Idrissi-Yaghir[4], Asma Ben Abacha[6], Alba G. Seco de Herrera[7], Henning Müller[8,9] and Christoph M. Friedrich[1,2]

[1]*Department of Computer Science, University of Applied Sciences and Arts Dortmund, Dortmund, Germany*
[2]*Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen, Germany*
[3]*Institute for Transfusion Medicine, University Hospital Essen, Essen, Germany*
[4]*Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen, Germany*
[5]*Data Integration Center, Central IT Department, University Hospital Essen, Essen, Germany*
[6]*Microsoft, Redmond, Washington, USA*
[7]*School of Computer Science, National University of Distance Education (UNED), Spain*
[8]*University of Applied Sciences Western Switzerland (HES-SO), Switzerland*
[9]*University of Geneva, Switzerland*

Abstract

The ImageCLEFmedical 2025 Caption task follows challenges held from 2017–2024 and comprises three subtasks: concept detection, caption prediction, and a newly introduced explainability task. The goal is to extract Unified Medical Language System (UMLS) concepts, generate fluent captions from medical images, and provide human-interpretable justifications for the outputs. This year's edition used an enlarged version of the Radiology Objects in COntext version 2 (ROCOv2) dataset, which was expanded with new articles and the inclusion of the optical coherence tomography (OCT) imaging modality. For concept detection, the F1-score was used to evaluate predictions against UMLS terms. For caption prediction, evaluation was updated to a composite score averaging six metrics to assess both relevance and factuality. The new explainability submissions were manually judged by a radiologist. The 2025 task attracted 80 registered research groups, with 11 teams submitting a total of 149 graded runs across the three subtasks. Top-performing systems for concept detection were predominantly based on ensembles of Convolutional Neural Networks (CNNs). For caption prediction, a general shift towards fine-tuning Vision-Language Models (VLMs) was observed, with adapted architectures like BLIP leading to strong results across the new composite metrics. Finally, the inaugural explainability task saw initial submissions of post-hoc visualizations, establishing a baseline and clarifying the need for model-intrinsic explanations in future editions.

Keywords

ImageCLEF, Computer Vision, Multi-Label Classification, Image Captioning, Image Understanding, Radiology, Explainable AI

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
†These authors contributed equally.
hendrik.damm@fh-dortmund.de (H. Damm); tabea.pakull@uk-essen.de (T. M. G. Pakull); helmut.becker@uk-essen.de (H. Becker); benjamin.bracke@fh-dortmund.de (B. Bracke); bahadir.eryilmaz@uk-essen.de (B. Eryılmaz); louise.bloch@fh-dortmund.de (L. Bloch); raphael.bruengel@fh-dortmund.de (R. Brüngel); cynthia.schmidt@uk-essen.de (C. S. Schmidt); johannes.rueckert@fh-dortmund.de (J. Rückert); obioma.pelka@uk-essen.de (O. Pelka); henning.schaefer@uk-essen.de (H. Schäfer); ahmad.idrissi-yaghir@uk-essen.de (A. Idrissi-Yaghir); abenabacha@microsoft.com (A. Ben Abacha); alba.garcia@lsi.uned.es (A. G. Seco de Herrera); henning.mueller@hevs.ch (H. Müller); christoph.friedrich@fh-dortmund.de (C. M. Friedrich)
0000-0002-7464-4293 (H. Damm); 0009-0009-9802-7167 (T. M. G. Pakull); 0000-0003-4986-7142 (B. Bracke); 0009-0002-8743-4751 (B. Eryılmaz); 0000-0001-7540-4980 (L. Bloch); 0000-0002-6046-4048 (R. Brüngel); 0000-0003-1994-0687 (C. S. Schmidt); 0000-0002-5038-5899 (J. Rückert); 0000-0002-4123-0406 (H. Schäfer); 0000-0003-1507-9690 (A. Idrissi-Yaghir); 0000-0001-6312-9387 (A. Ben Abacha); 0000-0002-6509-5325 (A. G. Seco de Herrera); 0000-0001-6800-9878 (H. Müller); 0000-0001-7906-0038 (C. M. Friedrich)

1. Introduction

ImageCLEF[1] [1] is the image–retrieval and –classification lab of the Conference and Labs of the Evaluation Forum (CLEF) conference [2]. ImageCLEF 2025 [3] consists of the ImageCLEFmedical, ImageCLEFrecommending, Image Retrieval for Arguments (Touché) and ImageCLEFToPicto labs, with the ImageCLEFmedical lab being divided into the subtasks Caption (image–captioning), VQA (text-to-image generation), MEDIQA-MAGIC (Multimodal And Generative TelemedICine) and GANs (generation of medical images).

The Caption task was first proposed as part of the ImageCLEFmedical [4] in 2016. In 2017 and 2018 [5, 6] it comprised two subtasks: concept detection and caption prediction. From 2019 [7] to 2020 [8] the focus shifted to concept detection, extracting Unified Medical Language System® (UMLS) [9] Concept Unique Identifiers (CUIs) from radiology images. Since 2021 [10] both subtasks have run in parallel again, with gradually higher-quality, manually annotated data and—in 2023—a switch from BLEU [11] to BERTScore [12] as the primary caption-prediction metric [13]. The 2024 edition introduced a small-scale explainability trial and an enlarged metric set.

2025 marks the 9th edition of the ImageCLEFmedical Caption task. Building on the lessons of previous years, the task now comprises three components:

1. Concept Detection – identification of UMLS concepts in radiology images;
2. Caption Prediction – generation of coherent captions for full images;
3. Explainability – newly promoted to an official subtask: participants must provide human-interpretable explanations for a designated subset of images, which are manually judged by a radiologist for interpretability, relevance and creativity.

For caption prediction, the overall ranking is now based on the average across these six metrics, reflecting both relevance and factuality aspects of the generated captions.

Manual creation of structured knowledge from medical images is slow and error-prone. By benchmarking automatic systems that detect clinical concepts, compose fluent radiology captions and justify their outputs, ImageCLEFmedical 2025 continues to stimulate research toward scalable, trustworthy radiology-image understanding.

As in 2024, the development data are drawn from an *extended* version of the Radiology Objects in COntext Version 2 (ROCOv2) dataset [14]. For 2025, this release has been enlarged with additional, newly released PubMed Central® Open-Access articles whose images and captions were again manually annotated with modalities. A novelty to this year's dataset is the inclusion of the imaging modality *optical coherence tomography (OCT)*, which has been retrospectively annotated for every existing ROCOv2 image and prospectively annotated for all new articles. The final split now comprises 80 091 training, 17 277 validation, and 19 267 test radiology images, all with updated licensing curation and UMLS (2022 AB) concept filtering.

This paper presents an overview of the ImageCLEFmedical 2025 Caption task: the task design and participation (Section 2), data creation (Section 3), evaluation methodology (Section 4), results (Section 5) and conclusions (Section 6). Further information on the other ImageCLEF 2025 tasks can be found in Ionescu et al. [3].

2. Task and Participation

For the 9th edition, the ImageCLEFmedical Caption task builds on two familiar subtasks:

- **T1 Concept Detection**. Systems predict Unified Medical Language System® (UMLS) Concept Unique Identifiers (CUIs) [9] directly from radiology images, following the format introduced in 2017 [5].

[1] https://www.imageclef.org/ [last accessed: 2025-06-01]

- **T2 Caption Prediction**. Systems generate full-sentence captions for each image, a subtask that returned in 2021 after a pause in 2019–2020.

and introduces a third, officially-graded component:

- **Exp Explainability**. For a small radiologist-selected subset, each team provides one human-interpretable explanation (for example a heat-map, bounding boxes or a textual rationale) that relates the image to the generated caption. This explanation is intended to clarify the model's decision-making process and thereby support clinicians in building trust in the model. Explanations are judged manually by a radiologist for interpretability, clinical relevance and creativity.

The 2025 edition also adds six evaluation metrics for caption prediction (see Section 4) and retrospectively annotates the complete ROCOv2 corpus with the new optical coherence tomography (OCT) modality. To compensate for the greater computational effort and occasional Docker-induced submission problems, the limit for graded runs per team was raised to **30** for T1 and T2; previously, it had been set at 10 runs. The Explainability Task (Exp) only allowed one submission, due to manual evaluation effort.

2.1. Participation Statistics

Eighty research groups signed the End-User Agreement and downloaded the development data. Eleven of them submitted runs and ten provided accompanying working-note papers. The submissions were distributed across the tasks as follows:

- Concept Detection (T1): 9 teams, 51 graded runs.
- Caption Prediction (T2): 8 teams, 98 graded runs.
- Explainability (Exp): 2 teams, 2 graded runs.
- Total: 149 graded runs.

Six groups took part in both T1 and T2. Three teams (DeepLens, mapan and LekshmiscopeVIT) focused on concept detection only, and two (CSMorgan and AI Stat Lab) entered just the caption-prediction track. Five teams, AUEB NLP Group, UIT-Oggy, CS_Morgan, sakthiii and LekshmiscopeVIT, had already participated in 2024 and are marked with an asterisk in Table 1.

The 2025 task therefore attracted a participant pool similar in size to earlier editions but generated more graded submissions, while also promoting explainability to a fully assessed subtask.

3. Data Creation

Figure 1 illustrates a typical sample from this year's collection. The following subsections describe the process of data collection, preprocessing, and annotation in detail, highlighting key decisions and challenges encountered during the creation of the dataset.

3.1. Source and Split

All data originate from articles in the PubMed Central® (PMC) Open-Access subset[2] [25]. The development data correspond to an *extended* release of ROCOv2 [14], enlarged with all papers published between October 2022 and December 2024. Captions were only stripped of URLs and non-English captions were dropped.

The final dataset is split into **80 091** training, **17 277** validation and **19 267** test images (**116 635** in total).

[2] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ [last accessed: 2025-06-01]

Table 1
Participating groups in the ImageCLEFmedical 2025 Caption task and their graded runs submitted to the three subtasks.

Team	Institution	Runs T1	Runs T2	Runs Exp
AUEB NLP Group* [15]	Athens University of Economics and Business, Greece	16	26	1
DeepLens* [16]	Iran University of Science and Technology, Tehran, Iran	14	–	–
mapan	-	4	–	–
UIT-Oggy* [17]	University of Information Technology, Ho Chi Minh City, Vietnam	8	23	–
DS4DH [18]	Hunan City University, China	1	11	–
sakthiii* [19]	Rajalakshmi Engineering College, Chennai, India	1	1	–
JJ-VMed [20]	Universidad Europea de Valencia, Spain	1	2	1
UMUTeam [21]	University of Murcia, Spain	2	2	–
LekshmiscopeVIT* [22]	Vellore Institute of Technology, Chennai, India	4	–	–
CSMorgan* [23]	Morgan State University, Baltimore, USA	–	5	–
AI Stat Lab [24]	Chung-Ang University, Seoul, Republic of Korea	–	28	–

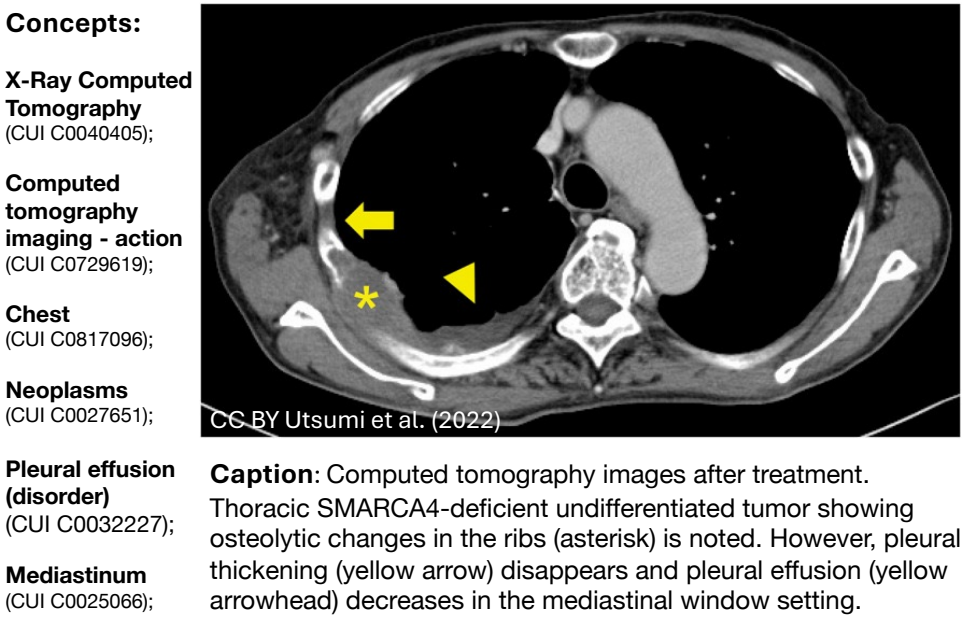

Figure 1: Example image, caption and UMLS® CUIs from the ImageCLEFmedical Caption 2025 task (CC-BY, Utsumi et al. 2022).

3.2. Concept Extraction

Concepts were extracted with MedCAT [26] trained on MIMIC-III [27] and mapped to UMLS 2022AB CUIs. Only concepts occurring at least ten times and belonging to semantically "visible" TUI groups were kept; ambiguous or spurious concepts were merged or removed through manual curation.

3.3. Modality and Region Concepts

Each image is manually labelled with an imaging-modality concept. In addition to the five modalities used in previous editions (X-ray, CT, MRI, ultrasound, PET/PET-CT) the 2025 corpus introduces **optical coherence tomography (OCT, CUI C0920367)**. OCT was annotated retrospectively for the entire

archive and prospectively for new articles.

Table 2 lists the modality distribution, while Table 3 details the image retrieval in medical applications (IRMA) region counts.

Table 2
Number of images per modality.

Modality	Images
CT	40 913
X-ray	31 827
MRI	18 570
Ultrasound	17 147
Angiography	6 055
PET	1 134
PET/CT	580
OCT	409

Table 3
Distribution of IRMA body-region concepts in X-ray images.

Region	CUI	Images
Chest	C0817096	10 931
Cranium	C0037303	5 436
Lower extremity (leg)	C0023216	4 563
Abdomen	C0000726	3 490
Upper extremity (arm)	C1140618	2 188
Pelvis	C0030797	1 923
Spine	C0037949	1 823
Other / unclear	—	1 145
Breast (mamma)	C0006141	210

3.4. Concept Statistics

Table 4 compares the concept inventory of the 2025 corpus with the preceding three editions. While the total image count has increased substantially, the number of unique concepts has grown only moderately. This reflects the effectiveness of concept pruning and semantic filtering in keeping the label space manageable.

3.5. Released Sets

- *Training set*: 80 091 images, 252 772 concept occurrences, 1 949 unique concepts.
- *Validation set*: 17 277 images, 48 761 concept occurrences, 716 unique concepts.
- *Test set*: 19 267 images, 24 242 concept occurrences, 702 unique concepts.
- *Explainability set*: 16 images (two from each modality, including two OCT cases) were selected by a radiologist based on the clinical relevance of both the images and their corresponding captions for manual assessment. In addition, examples of how such explanations might look like are provided, which can be found in Figure 2.

4. Evaluation Methodology

This year, the evaluation procedure was revised to reflect improved methodology and the incorporation of new tools and metrics. As in previous editions, the subtasks were evaluated independently.

Table 4
Unique concepts and average concepts per image by split for the ImageCLEFmedical Caption datasets 2022–2025.

Year	Split	Unique concepts	Concepts / image
2022	Train	17 210	4.90
	Valid	5 126	4.85
	Test	4 403	4.97
2023	Train	2 126	3.73
	Valid	1 946	3.84
	Test	1 936	3.86
2024	Train	1 946	3.15
	Valid	1 752	3.21
	Test	700	2.82
2025	Train	1 973	3.17
	Valid	716	2.83
	Test	702	3.06

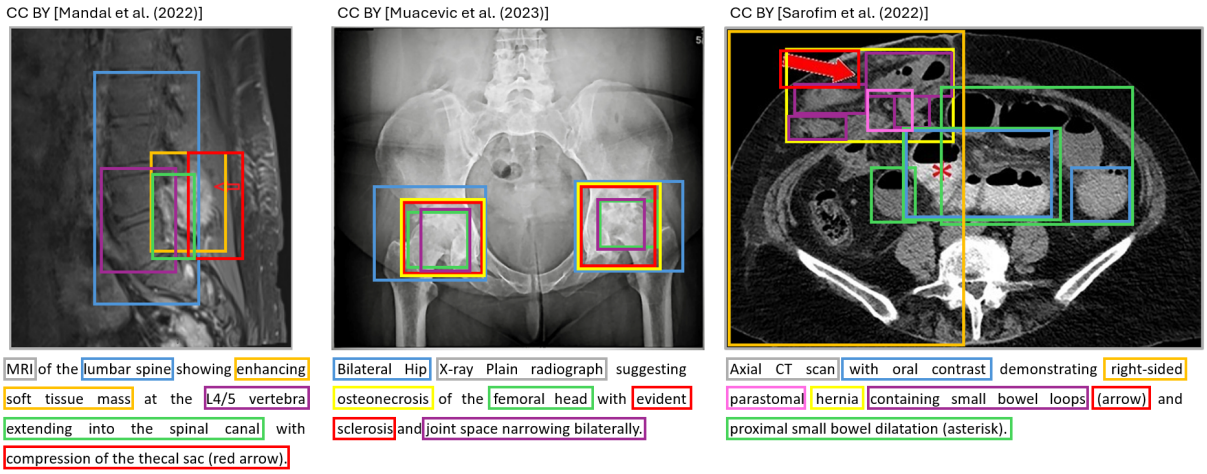

Figure 2: Example image, to demonstrate how explanations for captions might look like.

In 2025, the AI4MediaBench[3] by AIMultimediaLab[4] was used as the challenge platform.

For the concept detection subtask, the balanced precision and recall trade-off were measured in terms of F1-scores. Like last year, a secondary F1-score is computed using a subset of concepts that was manually curated. On the one hand, this involves the different image modalities (X-ray, Angiography, Ultrasound, CT, MRI, PET, OCT, and Combined such as PET/CT). On the other hand, if applicable, for X-ray also the anatomical code for body region examined of IRMA (cranium, chest, upper extremity, spine, abdomen, pelvis, and lower extremity) was involved.

For caption prediction, system outputs were assessed using a composite score, averaging across six complementary metrics to jointly capture aspects of relevance and factuality. All individual scores for each caption are summed and averaged over the number of captions, resulting in the final score.

Relevance was evaluated using four different methods. The first of these is BERTScore [12], which is a metric that computes a similarity score for each token in the generated text with each token in the reference text. It uses the pre-trained contextual embeddings from Bidirectional Encoder Representations from Transformers (BERT) [28]-based models and matches words by cosine similarity. In this work, the

[3]https://ai4media-bench.aimultimedialab.ro/ [last accessed: 2025-06-02]
[4]https://www.aimultimedialab.ro/ [last accessed: 2025-06-02]

pre-trained model *microsoft/deberta-xlarge-mnli*[5] was used because it is the model that correlates best with human scoring according to the authors[6]. Following best practices for caption evaluation reported by [12], we computed Recall-based BERTScore with inverse document frequency (idf) weighting, using idf scores derived from the test set to emphasize informative terms. The second metric, ROUGE (Recall-Oriented Understudy for Gisting Evaluation [29]) score, counts the number of overlapping units such as n-grams, word sequences, and word pairs between the generated text and the reference. Specifically, the ROUGE-1 (F-measure) score was calculated, which measures the number of matching unigrams between the model-generated text and a reference. The third relevance metric BLEURT (BiLingual Evaluation Understudy with Representations from Transformers) [30] is designed to assess the quality of natural language generation in English by leveraging a pre-trained model that has been fine-tuned to emulate human judgments about the quality of the generated text. The strength of BLEURT lies in its end-to-end training, which enables it to model human judgments effectively and makes it robust to domain and quality variations. For this evaluation, the BLEURT-20 model was used.

All of the above-mentioned metrics were computed using preprocessed captions that were lowercased and had punctuation stripped. Numeric values were replaced with the token "number." The captions were treated as single sentences, regardless of actual sentence boundaries. This step ensures uniformity and focuses the evaluation on linguistic content.

In addition to the text-based metrics a reference free metric was implemented. The methodology is based on CLIPScore [31], an innovative metric that diverges from the traditional reference-based evaluations of image captions. Instead, it aligns with the human approach of evaluating caption quality without references by evaluating the alignment between text and image content. The original metric employs Contrastive Language-Image Pretraining (CLIP) [32], a cross-modal model that has been pre-trained on a massive dataset of image-caption pairs sourced from the web. For this year's evaluation the MedImageInsight [33] model was used instead. It is trained using medical images with associated text and labels from a variety of domains, including X-ray, CT, MRI, OCT, and ultrasound. The model is used to compute similarity scores between images and text.

To assess the factuality of the generated captions, two complementary metrics were employed. The UMLS Concept F1-scoreevaluates the overlap of medical entities between the generated and reference captions. Specifically, medical concepts were extracted using MedCAT [34], with a focus on semantic types relevant to clinical accuracy as also defined for the MEDCON [35] metric whereas MEDCON relies on QuickUMLS [36] for concept extraction from both texts. This is followed by calculation of the F1-score to quantify concept-level agreement. The other factuality metric, AlignScore [37], employs a deep learning approach based on RoBERTa [38] to measure factual consistency. It involves the decomposition of extensive texts into more manageable segments and aligning the claims in the generated caption with the supporting evidence in the reference caption, thereby producing an average alignment score across all claims.

For the explainability extension, a radiologist was asked to rate both, the caption and the visualisation of each image in the explainability subset on a 1-5 Likert scale, with 5 being the best score.

The captions were ranked in terms of readability, clinical appropriateness, level of detail, and focus. The readability scale ranks whether the predicted captions are readable and coherently formulated. The clinical appropriateness evaluates whether the predicted captions match ground-truth captions or are clinically plausible. The level of detail is used to assess whether the captions merely describe visual findings or also interpret underlying clinical concepts. The focus validates the appropriateness of the scope of the caption and thus penalizes short captions that lack essential observations as well as excessively long captions that are not focused on the essentials.

The visualisation was assessed based on visual-text coherence, completeness, and focus. The visual-text coherence measures, if the visualisation is comprehensible in relation to the predicted caption. The completeness scale assesses, whether the visualisations meet all relevant concepts. The focus validates the appropriateness of the visualisation.

[5] https://huggingface.co/microsoft/deberta-xlarge-mnli [last accessed: 2025-06-05]
[6] https://github.com/Tiiiger/bert_score [last accessed: 2025-06-05]

Each image was rated individually and an average score across categories was reported. In addition, the radiologist rates the meaningfulness of the overall methodology. The final score was calculated as the average of all criteria.

5. Results

For the concept detection and caption prediction subtasks, Tables 5 and 6 show the best results from each of the participating teams. The results will be discussed in this section. The full list of results are shown in Appendix A in Tables 12, 13 and 15. Finally, Table 9 presents the results for the explainability subtask.

5.1. Results for the Concept Detection Subtask

In 2025, 9 teams participated in the concept prediction subtask, submitting 51 graded runs. Table 5 presents the best results for each team achieved in the submissions.

Table 5
Performance of the participating teams in the ImageCLEFmedical 2025 Caption concept detection subtask. Only the best run based on the achieved F1-score is listed for each team, together with the corresponding secondary F1-score based on manual annotations as well as the team rankings based on the primary and secondary F1-score. The full results are shown in Table 12 in Appendix A.

Group Name	Best Run	F1	Secondary F1	Rank (secondary)
AUEB NLP Group	1980	**0.5888**	**0.9484**	1 (1)
DeepLens	1725	0.5766	0.9299	2 (2)
mapan	1505	0.5660	0.9298	3 (3)
UIT-Oggy	1892	0.5613	0.9104	4 (4)
DS4DH	1508	0.5225	0.8672	5 (6)
sakthiii	1774	0.4003	0.9082	6 (5)
JJ-VMed	1903	0.3982	0.8329	7 (7)
UMUTeam	1807	0.2398	0.5377	8 (8)
LekshmiscopeVIT	1942	0.1494	0.2298	9 (9)

AUEB NLP Group [15] The AUEB NLP Group based their approach on their past work, which won the competition many years, but reached second place in the last year. The approach combined CNNs (EfficientNet-B0 [39], DenseNet-121 [40], and ConvNeXt-Tiny [41]) with per-label threshold optimization and ensembling strategies, including dual threshold aggregation, and partial intersection aggregation. The team won the first place with a primary F1-score of 0.5888 and a secondary F1-score of 0.9484.

DeepLens [16] The DeepLens team tackled the concept detection task with an ensemble model pipeline which combined EfficientNet-B0 [39] and DenseNet-121 [40] under a simple union ensemble. Both networks were optimized with the ADAM optimizer using the Binary Cross Entropy with Logits loss function. The output layers of the models were replaced either with a three-layer feed-forward head or a single linear classifier to finetune the models for multi-label prediction. The ensemble with the best micro-F1-scorevalidation score was frozen for test inference. This method delivered the team's best submission, securing a primary F1-scoreof 0.5766 and a secondary F1-scoreof 0.9299, which placed second overall in the competition. Furthermore, the DeepLeans team experimented with a K-Nearest Concept-Language-Image Pre-training to improve image-concept alignment in their ensemble strategy. Although it did not yield the best quantitative results, it might hold interesting directions for future research.

UIT-Oggy [17] For the concept detection task, the team designed MedCSRA, a novel architecture featuring a dual-branch design that combines global semantic understanding through global

average pooling with localized class-specific residual attention (CSRA) mechanisms. Four CNN backbones were evaluated: ResNet-101, DenseNet121, EfficientNet-B4 and EfficientNet-B5. All were pre-trained on ImageNet and fine-tuned for medical multi-label classification using Binary Cross Entropy Loss. The final prediction uses a weighted combination of the outputs from the global and CSRA branches. ResNet-101 achieved the highest F1-score of 0.5613, demonstrating that specialized attention mechanisms can effectively identify multiple medical concepts in biomedical images.

DS4DH [18] reformulated concept detection as an image-to-sequence task to leverage transformer-based models capable of capturing the inherent order of UMLS codes (e.g., modality before anatomy or pathology). They proposed a compact architecture combining a convolutional neural network to extract low-dimensional image embeddings (as small as 16 dimensions) with a lightweight transformer decoder (1 head, 2 layers) that autoregressively generates UMLS code sequences via cross-attention. Beam search (width = 3) was used during decoding and improved performance. This approach achieved an F1-score of 0.5225 and a secondary F1-score of 0.8672, ranking the team fifth and sixth, respectively. To address class imbalance, the team experimented with focal loss, label smoothing, and pre-trained embeddings (MedCPT [42], CUI2Vec [43]), but none outperformed their baseline model.

They observed that their model tended to produce short sequences (average length 1.3 CUIs) with low diversity (15 unique predicted CUIs), which they attributed to dataset bias toward short and imbalanced annotations. Applying loss masking strategies during training increased the average sequence length to 3.0 CUIs and raised diversity to 103 unique CUIs. However, this revised model underperformed in terms of F1-score compared to their baseline submission. The team suggested this discrepancy may result from the challenge's F1-score evaluation design, which potentially favors shorter CUI sequences and penalizes longer, yet possibly correct predictions not aligned with the ground-truth test data.

sakthiii [19] For the concept detection task, team sakthiii employed a MedCLIP-based transformer model, which was pre-trained on medical image-caption pairs. In the first stage of their dual-stage training pipeline, they fine-tuned this MedCLIP model specifically for concept detection. This process involved training for 11 epochs with a batch size of 32, using the Adam optimizer and a learning rate of 1e-5. The dataset for this stage consisted of radiology images paired with UMLS concepts, allowing the model to learn the mappings between visual features and structured medical terms. Their best model for concept detection achieved an F1-score of 0.4003 and a secondary F1-score of 0.9082 , placing them eighth in this subtask.

JJ-VMed [20] The JJ-VMed team employed a fine-tuned LLaVA-LLaMA 3 8B model, processing inputs through a CLIP ViT-Large encoder. Training used prompt-based instruction tuning, and two output formats were explored: one generating concepts independent from the caption, while the second embedded them within full-text captions. They achieved a primary F1-score of 0.3982 and a secondary F1-score of 0.8329, ranking them seventh in this subtask.

UMUTeam [21] Based on the captions generated by a fine-tuned BLIP model, the UMUTeam employed named entity recognition (SciSpacy), concept retrieval (SapBERT), followed by a BERT-based reranking classifier, to extract the medical concepts for the concept detection subtask. They achieved an F1-score of 0.2398 with a secondary F1-score of 0.5377, putting them in eighth place, showing that this caption-based approach is inferior to multi-label classification systems.

LekshmiscopeVIT [22] Team LekshmiscopeVIT focused on a broader evaluation of different deep learning architectures to approach the concept detection subtask. The team employed the standard architectures InceptionV3, DenseNet, and ResNet as well as a custom approach. Randomly initialized and ImageNet [44] pre-trained models of each of the standard architectures were fine-tuned on the ROCOv2 dataset for 10 epochs and then compared. Part of each training

pipeline was a uniform pre-processing step during which a multi-label binarizer was applied to create a binary label matrix for training. The team further experimented with reduction of label space complexity by limiting predictions to the most frequent concepts. The pre-trained ResNet approach achieved the team's best results of 0.1494 in the primary, and 0.2298 in the secondary F1-score.

The Concept Detection task this year revealed several methodological trends among the participating teams. The top-performing approaches relied on convolutional neural network (CNN) ensembles, combining multiple pre-trained architectures, such as EfficientNet, DenseNet, and ResNet. These ensembles used fine-tuned classification heads and per-label threshold optimization to improve multi-label prediction accuracy. Both simple and complex ensembling techniques proved effective, suggesting that leveraging the complementary strengths of different models remains strong.

Although CNNs dominated the leaderboard, several teams explored transformer-based and generative approaches. These included image-to-sequence formulations and vision-language models, such as MedCLIP and LLaVA. Though these methods were less competitive in terms of F1-scores, they indicate a growing interest in multimodal models.

Lower-ranking submissions often relied on caption-based pipelines and traditional CNNs without extensive optimization or innovative architectures. These underperformed compared to more tailored solutions.

A comparison of the 2024 and 2025 ImageCLEFmedical Concept Detection subtasks reveals a decline in primary F1-scores across the leaderboard, suggesting that this year's task may have been more challenging or less suited to the models deployed.

Despite this overall decline in primary performance, secondary F1-scores based on manual annotations remained high and in some cases even improved. For example, the AUEB NLP Group, which participated in both years, saw a drop in primary F1-score, but an increase in secondary F1-score from 0.9393 to 0.9484, reclaiming the top spot.

By training and evaluating our own baseline model on the data from this year, we could determine that about 0.1 of the difference in primary F1-score is purely due to the new test dataset, which contains a much smaller number of unique concepts (see Table 4).

The observed decline in primary F1-scores can likely be attributed to several interrelated factors stemming from changes in the dataset. First, the slight increase in average concepts per image introduced greater multi-label complexity, making it more difficult to make fully correct predictions under the strict F1-score metric. Second, the broader inclusion of imaging modalities, particularly the addition of optical coherence tomography (OCT) and expanded angiography cases, may have introduced domain shifts that negatively affected models that were not trained or tuned on such data. Lastly, although concept filtering improved label quality, it may have also limited the label space, penalizing over-predictive or less conservative systems.

5.2. Results for the Caption Prediction Subtask

In this edition, the caption prediction subtask attracted 8 teams which submitted 98 graded runs. Tables 6, 7 and 8 present the results of the submissions.

UMUTeam [21] The UMUTeam employed the BLIP [45] architecture, which consists of a ViT encoder and a language model decoder, to generate captions for medical images. They fine-tuned a model which performs well in general image captioning benchmarks, selecting the best model based on the relevance metric. With a score of 0.9271 for Similarity, 0.5977 for BERTScore Recall, 0.2594 for ROUGE-1, 0.3230 for BLEURT and an overall score of 0.3432, they won the caption prediction subtask, scoring highest in all but the BERTScore Recall and AlignScore metrics.

DS4DH [18] developed multiple strategies for automatic medical image captioning. First, they fine-tuned a Vision-Language Model (InstructBLIP-Flan-T5-XL [46]) using selective parameter freezing,

Table 6
Performance of the participating teams in the ImageCLEFmedical 2025 Caption caption prediction subtask. Only the best run based on the achieved Overall Score is listed for each team as well as the team rankings based on Overall Score together with rankings based on Relevance (Rel.) and Factuality (Fact.) Average. Additional scores are shown in Tables 7 and 8. The full results are shown in Tables 13 and 15 in Appendix A.

Group Name	Best Run	Overall	Relevance	Factuality	Rank (Rel./Fact.)
UMUTeam	1681	**0.3432**	**0.5268**	**0.1596**	1 (1/1)
DS4DH	1520	0.3362	0.5174	0.1549	2 (2/2)
AI Stat Lab	1900	0.3229	0.5089	0.1369	3 (3/3)
UIT-Oggy	1914	0.3211	0.5076	0.1346	4 (4/4)
AUEB NLP Group	1403	0.3068	0.4759	0.1377	5 (6/5)
JJ-VMed	1896	0.3043	0.4922	0.1165	6 (5/6)
sakthiii	1890	0.2746	0.4481	0.1011	7 (7/7)
CS_Morgan	1815	0.2315	0.3717	0.0917	8 (8/8)
Baseline (Llama 4 Scout)		0.3101	0.5073	0.1128	

Table 7
Performance of the participating teams in the ImageCLEFmedical 2025 Caption caption Prediction subtask for relevance metrics Similarity, BERTScore (Recall), ROUGE-1 and BLEURT. These correspond to the best Overall-based runs of each team, listed in Table 6. The full results are shown in Tables 13 and 15 in Appendix A.

Group Name	Best Run	Similarity	BERTScore (Recall)	ROUGE-1	BLEURT
UMUTeam	1681	**0.9271**	0.5977	**0.2594**	**0.3230**
DS4DH	1520	0.9016	**0.6067**	0.2516	0.3096
AI Stat Lab	1900	0.8919	0.5823	0.2440	0.3173
UIT-Oggy	1914	0.8798	0.5951	0.2535	0.3020
AUEB NLP Group	1403	0.7947	0.5884	0.2176	0.3030
JJ-VMed	1896	0.8251	0.5953	0.2389	0.3094
sakthiii	1890	0.7957	0.5553	0.1607	0.2806
CS_Morgan	1815	0.5704	0.5180	0.1598	0.2385
Baseline (Llama 4 Scout)		0.9360	0.5598	0.2078	0.3258

Table 8
Performance of the participating teams in the ImageCLEFmedical 2025 Caption caption Prediction subtask for factuality metrics UMLS Concept F1-score and AlignScore. These correspond to the best Overall-based runs of each team, listed in Table 6. The full results are shown in Tables 13 and 15 in Appendix A.

Group Name	Best Run	UMLS Concept F1	AlignScore
UMUTeam	1681	**0.1816**	0.1375
DS4DH	1520	0.1682	**0.1417**
AI Stat Lab	1900	0.1524	0.1213
UIT-Oggy	1914	0.1672	0.1021
AUEB NLP Group	1403	0.1429	0.1325
JJ-VMed	1896	0.1366	0.0964
sakthiii	1890	0.1094	0.0928
CS_Morgan	1815	0.0741	0.1087
Baseline (Llama 4 Scout)		0.1302	0.0955

focusing training on cross-modal alignment while keeping most of the vision and language encoders fixed. Second, they implemented a Retrieval-Augmented Generation [47] (RAG) approach

that retrieves visually similar training images and incorporates their captions into the prompt to guide caption generation. Third, they introduced a Cluster-based RAG strategy that groups training data by the semantic similarity of CUI codes using MedCPT [42] embeddings, enabling hierarchical retrieval within medically relevant clusters. Finally, they trained an alignment model (BioBart-v2-large [48]) using pairs of InstructBLIP-generated and ground-truth captions to refine caption quality.

Among all approaches, the fine-tuned InstructBLIP model achieved the highest overall score (0.3708) and ranked first in the recall-based BERTScore metric (0.6067) among all challenge participants. In contrast, both the alignment model and standard RAG approach underperformed, likely due to the introduction of noisy or irrelevant information, which reflects the visual similarity but semantic variability of radiology images. The Cluster-based RAG showed moderate improvements over standard RAG (e.g., overall score improved from 0.3478 to 0.3620). However, due to possible noise in predicted CUIs (F1-score= 0.5225) from the concept detection subtask, it still fell short of InstructBLIP. On the validation dataset, Cluster RAG outperformed InstructBLIP on several metrics when ground-truth CUIs were used. This highlights the critical importance of accurate concept detection for precise RAG retrieval cues, because even minor inaccuracies in CUI prediction can introduce semantic noise and significantly degrade caption quality.

AI Stat Lab [24] The team developed a modular framework for medical image captioning that begins with a two-stage preprocessing pipeline. This includes 2× super-resolution and inpainting to eliminate bright border artifacts. A dual-encoder setup (SigLIP2 [49] + BioMedCLIP [50]) feeds into a Q-Former [51], which generates concept-aware tokens used for both captioning and medical concept classification. A LoRA-tuned [52] Bio-Medical LLaMA-3-8B [53] serves as the decoder. Six model variants produce captions that are either summarized using GPT-4 [54] or reranked using custom-designed metrics: BioMedCLIP image-text alignment, BLEURT self-consensus, and BioBERT [55] centroid proximity. Their best submission used BioMedCLIP alignment, achieved an overall score of 0.3229 and ranked third overall.

UIT-Oggy [17] For this task, the UIT-Oggy team fine-tuned the BLIP model by using Vision Transformer (ViT) to encode images and BERT-based text decoding to generate medical captions. Images were preprocessed to a uniform resolution of 224×224 and captions were tokenised to a maximum length of 200 tokens, ensuring compatibility with the vision-language model's input requirements. The BLIP model achieved an overall score of 0.3211 for captioning, demonstrating the effectiveness of vision-language pre-training in adapting to the terminology and context of the medical domain.

AUEB NLP Group [15] The AUEB NLP Group's approach on caption prediction involved seven primary systems: A finetuned InstructBLIP [46] model, was extended by a synthesizer and multi-synthesizer approach, an LM-Fuser, and an Distance from Median Maximum Concept Similarity (DMMCS) mechanism. In addition a test-time-reranker based on MedCLIP [56] and a reinforcement learning-based Mixer were implemented. The team's best results were reached for the finetuned InstructBLIP model, which reached an overall rating of 0.3068 and the fifth rank in the challenge.

JJ-VMed [20] In the caption prediction task, JJ-VMed reused their LLaVA-LLaMA 3 model for initial generation, followed by post-processing with LLaMA 3.1. With a score of 0.8251 for Similarity, 0.5953 for BERTScore Recall, 0.2389 for ROUGE-1, 0.3094 for BLEURT and an overall score of 0.3043, they ranked sixth place in the caption prediction subtask.

sakthiii [19] Following the concept detection training, the team transitioned to the caption prediction task by reusing the same MedCLIP model weights. This second stage aimed to leverage the semantic understanding gained during concept identification to help generate contextually relevant textual descriptions for the images. For this task, each image was preprocessed, converted to

RGB format, and then paired with its corresponding caption from the dataset. The MedCLIP processor and tokenization pipeline from the Transformers library were utilized to prepare these multimodal inputs for the model. In the caption prediction task, their approach yielded scores of 0.7957 for Similarity, 0.5553 for BERTScore Recall, 0.1607 for ROUGE-1, and 0.2806 for BLEURT , also resulting in an eighth-rank achievement.

CS_Morgan [23] The CS_Morgan team investigated six distinct captioning pipelines by fine-tuning three vision-language backbones—Qwen-2B, Qwen2.5-3B, and SmolVLM-500M on the ROCOv2 dataset. They evaluated a vanilla LoRA-based adaptation (Submissions 1–3) and a modality-conditioned variant (Submissions 4–6) in which a ResNet-50 classifier (trained from scratch on four modalities: CT, MRI, Ultrasound, Radiograph) first predicts the image modality. During inference, the predicted modality label is concatenated to the prompt (e.g., "CT image: [image]. Describe the medical image.") to guide the caption generator toward modality-specific terminology. Across these six runs, Qwen-2B achieved the highest Overall score (0.2537) when fine-tuned without classification, while both Qwen2.5-3B and SmolVLM demonstrated improved BLEURT and MedCATs scores under modality-conditioned prompting. This two-stage pipeline highlights that even smaller models like SmolVLM-500M can approach mid-scale performance when provided with structured modality cues.

Baseline For this year's baseline models in the caption prediction subtask, we utilized off-the-shelf vision-language models to generate appropriate captions based on the challenge images. Specifically, we evaluated the performance of the following instruction-tuned models: Meta's LLaMA 4 Scout (17Bx16E) Instruct [57], Google DeepMind's Gemma 3 27B Instruct [58], and Alibaba Cloud's Qwen2.5-VL 32B Instruct [59]. Each model was prompted individually with the challenge images and the following standardized instruction prompt in-context:

> "You are a medical expert contributing to a peer-reviewed scientific journal. Your task is to write a caption for a medical image, exactly as it would appear beneath a figure in a PubMed-indexed article. Concisely describe the clinical content of the image, identifying the imaging modality, key medical concepts, anatomical structures, visible markings, and any relevant abnormalities or pathologies. Where appropriate, include standard abbreviations in addition to full terms for modality, medical concepts, and pathologies (e.g., 'magnetic resonance imaging (MRI)'). Do not include any explanations, introductions, titles, figure numbers (e.g., 'Figure 1:' / 'Fig 1:'), references, or bullet points. Text only the caption."

To ensure reproducibility, we employed a deterministic decoding strategy by setting the Top-k sampling parameter to $k = 1$, thereby always selecting the most likely predicted token at each step. Among the three baseline models evaluated, Meta's LLaMA 4 Scout (17Bx16E) Instruct model performed best, obtaining an overall challenge score of 0.3101. This result positioned it approximately in the middle range of the submitted participant approaches. Notably, LLaMA 4 Scout achieved the highest scores in the Similarity metric (0.9369) and BLEURT metric (0.3258).

In the 2025 ImageCLEFmedical Caption Prediction subtask, all participating teams used vision-language models (VLMs) as the basis for their methods, showing a clear trend of using recent advances in multimodal architectures. Most submissions used or fine-tuned Transformer-based models, such as BLIP, InstructBLIP, and LLaMA variants. This indicates a reliance on pretrained models with strong image-text alignment capabilities. Several teams incorporated retrieval-augmented generation (RAG), multi-stage pipelines, or modular architectures to improve alignment with medical content. However, performance gains from these methods varied depending on the accuracy of supporting components, such as concept detection systems. Additionally, some teams used post-processing strategies, such as reranking or summarization. Despite the variety of approaches, models with direct fine-tuning on medical data and minimal architectural complexity often outperformed more elaborate pipelines. This result highlights the continued relevance of focused adaptation.

The results of the ImageCLEFmedical 2025 Caption Prediction subtask indicate a notable shift in evaluation priorities from general linguistic similarity toward a more balanced assessment of relevance and clinical factuality. Teams such as UMUTeam and DS4DH exhibited strong performance across both the relevance and factuality dimensions, outperforming several returning participants.

The analysis indicates that linguistic similarity metrics, such as BERTScore and ROUGE, demonstrate a high degree of consistency with those observed in the previous year, suggesting stable performance in terms of surface-level textual alignment. Embedding-based similarity scores are notably elevated among the top-performing submissions, suggesting that the generated captions may encompass semantically relevant content that extends beyond the scope of the original reference captions. This finding suggests a potential discrepancy between lexical overlap and underlying semantic alignment. Factuality-oriented metrics such as UMLS Concept F1-score and AlignScore remain relatively low, underscoring the inherent difficulty of ensuring clinical accuracy in generated captions. However, reliance on the original captions as the sole reference may limit the effectiveness of these scores in evaluating the full range of medically plausible outputs.

5.3. Results for the Explainability Subtask

This year, two teams participated in the explainability subtask. Table 9 presents the summarised results for both teams. In addition,

Table 9
Performance of the participating teams in the ImageCLEFmedical 2025 Caption caption prediction explainability extension.

Group Name	Overall	Caption	Visualization	Methodology
AUEB-NLP-Group	3.2	3.3	2.8	4.0
JJ-VMed	2.6	3.2	1.9	2.0

Table 10 presents the results of the caption sub-scale and

Table 10
Performance of the participating teams in the ImageCLEFmedical 2025 Caption caption prediction explainability extension, caption subscales.

Group Name	Average score	Readability	Clinical appropriateness	Level of detail	Focus
AUEB-NLP-Group	3.3	4.5	2.7	2.6	3.3
JJ-VMed	2.6	3.4	2.4	2.8	4.1

Table 11 those of the visualisation sub-scale.

Table 11
Performance of the participating teams in the ImageCLEFmedical 2025 Caption caption prediction explainability extension, visualization subscales.

Group Name	Average score	Visual-text coherence	Completeness	Focus
AUEB-NLP-Group	2.8	3.1	2.8	2.6
JJ-VMed	1.9	1.9	1.9	1.9

AUEB NLP Group [15] The AUEB NLP Group extracted UMLS concepts of the captions generated by their finetuned InstructBLIP [46] model using a biomedical NER model of the ScispaCy library. GPT-4o was used to identify bounding boxes for these concepts. The group reached the best overall rating of 3.2 by the radiologist. However, it should be noted that the explainability

approach focuses solely on the generated captions and does not involve the black-box model itself, which means it does not enhance the radiologist's trust in the model's predictions.

JJ-VMed [20] For the explainability task, JJ-VMed implemented a three-phase approach: Spatial mapping using GPT-4 and GPT-4V to link concepts and textual descriptions with image regions, segmentation and object detection using SAM [60] (Segment Anything Model) and YOLOv8 [61], as well as visualisation heuristics, such as arrow-following and keypoint-detection. The outputs included bounding boxes, segmentation masks, and heatmaps. The team achieved an overall rating of 2.6. Similar to the winning approach, this method does not incorporate the black-box model itself, and therefore the explanations do not contribute to increasing trust in the model's predictions.

In summary, both approaches used bounding boxes to visualise the connection between the images and specific concepts of the captions. The JJ-VMed team also provided heatmaps. Both visualisation methods are clinically valid. Although similar visualisation methods were used, the underlying techniques used for generation strongly differed. While the AUEB NLP group combined NER with GPT-4o to generate bounding boxes, the JJ-VMed combined GPT-4V models with YOLO object detection and Segment anything models (SAM) for segmentation. Both of these methods used to generate the explainability visualisations are based on external models. These models have no direct integration with the black-box model responsible for generating the captions. In conclusion, the visualizations do not contribute to increase the clinicians' trust in the presented captioning model. More appropriate approaches for this task would be to use attention maps [62], GradCAM [63], or Layer-wise Relevance Propagation (LRP) [64], to generate model-intrinsic explanations that highlight the regions or features within the image that actually influenced the captioning output, thereby providing more meaningful insights into the model's decision-making process.

During the manual validation, it was found that both participating teams were generally able to identify the imaging modality and the approximate anatomical region depicted in the images. However, substantial limitations were observed in the accurate identification and spatial localization of anatomical structures and pathological findings. A recurring issue across both submissions involved the inaccurate placement, scale, and labeling of bounding boxes. Frequently, the annotations only partially covered the target anatomical entities or failed to capture them entirely. Both teams generated syntactically coherent and clinically plausible captions, though with notable differences in level of detail and accuracy. The AUEB NLP Group demonstrated greater accuracy in the identification and localisation of anatomical entities, resulting in more precise but less informative annotations. In contrast, JJ-VMed produced more detailed and descriptive captions, albeit often based on incorrect concept detection.

6. Conclusion

The 9th edition of the ImageCLEFmedical Caption task continued its evolution with three components: the established Concept Detection and Caption Prediction subtasks, and the promotion of Explainability to a fully graded subtask. This year's challenge introduced an enlarged dataset featuring the new Optical Coherence Tomography (OCT) modality and a revised evaluation framework for captioning. The task attracted 11 teams who submitted a total of 149 graded runs, a substantial increase in submissions fostered by a higher run quota. Participation was balanced, with six teams entering both core subtasks, three focusing solely on concept detection, and two on caption prediction. Two teams took on the new explainability challenge.

For the concept detection subtask, the top-performing methods continued to rely on powerful ensembles of Convolutional Neural Networks (CNNs). However, a notable trend was the exploration of transformer-based and generative approaches by several teams, signalling a potential shift in methodology for future challenges.

In the caption prediction subtask, a clear consensus emerged around vision-language models (VLMs), with all teams leveraging architectures like BLIP, LLaMA, and their variants. Interestingly, direct

fine-tuning on medical data often outperformed more elaborate pipelines, such as Retrieval-Augmented Generation (RAG), which proved sensitive to the quality of their retrieval components, highlighting the challenge of system interdependencies.

In a reversal from 2024, primary F1-scores for concept detection saw a general decline across the leaderboard. This is attributed to the increased difficulty of the 2025 dataset, which featured new modalities like OCT and greater multi-label complexity. Despite this, secondary F1-scores on curated concepts remained high, indicating that models still perform robustly on core clinical findings.

The introduction of a composite score for caption prediction, averaging six metrics for relevance and factuality, successfully shifted the focus toward a more holistic evaluation. While relevance scores were strong, factuality metrics like UMLS F1-scoreand AlignScore remain modest across all submissions, underscoring that generating clinically accurate text is still the primary hurdle for the field. Notably, an off-the-shelf LLaMA 4 Scout baseline proved competitive, establishing a strong benchmark and demonstrating that while large foundation models are powerful, specialised fine-tuning still provides a winning edge.

Looking ahead, a primary focus for the 2026 challenge will be on advancing the maturity of the explainability task. This year's initial submissions relied on post-hoc visualisations generated by external models. While a valid first step, these methods do not offer insights into the captioning model's internal decision-making process. Future iterations will therefore strongly encourage the development of model-intrinsic explanations, such as attention maps or GradCAM, to foster genuine trust in the underlying VLM. Furthermore, the 2026 edition will broaden the task's scope and realism. The dataset will be extended again with recent PubMed Central publications, and to address the multilinguality of scientific literature, non-English captions will be translated and incorporated into the dataset, whereas previously they were omitted. For images that lack a direct caption, a baseline description will be generated for the dataset by using the context from the source article. The introduction of multilingual data and a continued focus on model transparency are intended to stimulate further research toward capable and reliable medical image understanding systems.

Acknowledgments

The work of Louise Bloch, Benjamin Bracke 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 Ahmad Idrissi-Yaghir, Henning Schäfer, Tabea M. G. Pakull, Hendrik Damm, Helmut Becker, 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). This work was partly supported by the project GRESEL-UNED PID2023-151280OB-C22 funded by MICIU/AEI/ AEI 501100011033.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using these services, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] H. Müller, J. Kalpathy-Cramer, A. García Seco de Herrera, Experiences from the ImageCLEF Medical Retrieval and Annotation Tasks, Springer International Publishing, Cham, 2019, pp. 231–250. URL: https://doi.org/10.1007/978-3-030-22948-1_10. doi:10.1007/978-3-030-22948-1_10.

[2] N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF, Springer, Cham, 2019. URL: https://link.springer.com/book/10.1007/978-3-030-22948-1. doi:10.1007/978-3-030-22948-1.

[3] B. Ionescu, H. Müller, D.-C. Stanciu, A.-G. Andrei, A. Radzhabov, Y. Prokopchuk, Ştefan, Liviu-Daniel, M.-G. Constantin, M. Dogariu, V. Kovalev, H. Damm, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, B. Bracke, O. Pelka, B. Eryilmaz, H. Becker, W.-W. Yim, N. Codella, R. A. Novoa, J. Malvehy, D. Dimitrov, R. J. Das, Z. Xie, H. M. Shan, P. Nakov, I. Koychev, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, D. Fabre, C. Macaire, B. Lecouteux, D. Schwab, M. Potthast, M. Heinrich, J. Kiesel, M. Wolter, B. Stein, Overview of ImageCLEF 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.

[4] A. García Seco de Herrera, R. Schaer, S. Bromuri, H. Müller, Overview of the ImageCLEF 2016 medical task, in: Working Notes of CLEF 2016 (Cross Language Evaluation Forum), 2016, pp. 219–232.

[5] C. Eickhoff, I. Schwall, A. García Seco de Herrera, H. Müller, Overview of ImageCLEFcaption 2017 - image caption prediction and concept detection for biomedical images, in: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017., 2017. URL: http://ceur-ws.org/Vol-1866/invited_paper_7.pdf.

[6] A. García Seco de Herrera, C. Eickhoff, V. Andrearczyk, H. Müller, Overview of the ImageCLEF 2018 caption prediction tasks, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018., 2018. URL: http://ceur-ws.org/Vol-2125/invited_paper_4.pdf.

[7] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the ImageCLEFmed 2019 concept detection task, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/paper_245.pdf.

[8] O. Pelka, C. M. Friedrich, A. García Seco de Herrera, H. Müller, Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding, in: CLEF2020 Working Notes, volume 1166 of *CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece, 2020.

[9] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) 267–270. doi:10.1093/nar/gkh061.

[10] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021, pp. 1101–1112.

[11] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[13] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2023 – caption prediction and concept detection, in: CLEF2023 Working Notes, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, Thessaloniki, Greece, 2023, pp. 1328–1346.

[14] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. Ben Abacha, A. García Seco de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology Objects in COntext version 2, an updated multimodal image dataset, Scientific Data 11 (2024) 688. doi:10.1038/s41597-024-03496-6.

[15] A. Chatzipapadopoulou, I. Pantelidis, F. Charalampakos, M. Samprovalaki, G. Moschovis, P. Kaliosis, K. Dalakleidi, J. Pavlopoulos, I. Androutsopoulos, AUEB NLP group at ImageCLEFmedical Caption 2025, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain,

2025.
[16] A. H. S. Rudsari, B. K. Nejad, M. Hajihosseini, S. Eetemadi, Detecting concepts for medical images: Contributions of the DeepLens team at IUST to ImageCLEFmedical caption 2025, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[17] M. V. Luong, M. H. Dinh-Doan, G. P. Bui-Hoang, T. B. Nguyen-Tat, UIT-Oggy at ImageCLEFmedical 2024 caption: CSRA-enhanced concept detection and BLIP-driven vision-language captioning, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[18] J. He, S. Ferdowsi, W. Feng, F. Alves, A. Platon, D. Teodoro, DS4DH group at ImageCLEFmedical caption 2025, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[19] T. Sakthi Mukesh, A. Beulah, R. Muthulakshmi, ImageCLEF-medical 2025: MedCLIP model for medical caption prediction and concept detection, in: Working Notes of the Conference and Labs of the CLEF Association (CLEF 2025), Madrid, Spain, 2025. Notebook for the ImageCLEF Lab at CLEF 2025.
[20] J. Angulo, J. Aguilar, JJ-VMed: A framework for automated concepts, captions and explainability of medical image, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[21] R. Pan, T. Bernal Beltrán, J. A. García Díaz, R. Valencia-García, 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, 2025.
[22] A. Sahni, R. Gupta, R. Venugopal Reddy, L. Kalinathan, Evaluating deep CNNs for multi-label concept detection in ROCOv2 radiology image dataset by team LekshmiscopeVIT, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[23] R. N. Chowdhury, M. Hoque, M. R. Hasan, E. P. O. Oluwafemi, M. M. Rahman, Modality-guided radiology caption prediction with small vision-language models and image classifier, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.
[24] Y. Lee, H. J. Kim, H. Shin, C. Lim, A modular framework for clinically accurate medical image captioning using vision-language models, in: Working Notes of the Conference and Labs of the CLEF Association (CLEF 2025), Madrid, Spain, 2025. Notebook for the ImageCLEF Lab at CLEF 2025.
[25] R. J. Roberts, PubMed Central: The GenBank of the published literature, Proceedings of the National Academy of Sciences of the United States of America 98 (2001) 381–382. doi:10.1073/pnas.98.2.381.
[26] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, R. J. Dobson, Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit, Artificial Intelligence in Medicine 117 (2021) 102083. doi:10.1016/j.artmed.2021.102083.
[27] A. E. Johnson, T. J. Pollard, L. Shen, L. wei H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016). doi:10.1038/sdata.2016.35.
[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
[29] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.
[30] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for

Computational Linguistics, Online, 2020, pp. 7881–7892. doi:10.18653/v1/2020.acl-main.704.

[31] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7514–7528. doi:10.18653/v1/2021.emnlp-main.595.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[33] N. C. F. Codella, Y. Jin, S. Jain, Y. Gu, H. H. Lee, A. Ben Abacha, A. Santamaria-Pang, W. Guyman, N. Sangani, S. Zhang, H. Poon, S. Hyland, S. Bannur, J. Alvarez-Valle, X. Li, J. Garrett, A. McMillan, G. Rajguru, M. Maddi, N. Vijayrania, R. Bhimai, N. Mecklenburg, R. Jain, D. Holstein, N. Gaur, V. Aski, J.-N. Hwang, T. Lin, I. Tarapov, M. Lungren, M. Wei, MedImageInsight: An open-source embedding model for general domain medical imaging, 2024. URL: https://arxiv.org/abs/2410.06542. arXiv:2410.06542.

[34] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, R. J. B. Dobson, Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit, Artificial Intelligence in Medicine 117 (2021) 102083. doi:10.1016/j.artmed.2021.102083.

[35] W.-w. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, M. Yetisgen, Aci-bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation, Scientific Data 10 (2023) 586. doi:10.1038/s41597-023-02487-3.

[36] L. Soldaini, N. Goharian, QuickUMLS: A fast, unsupervised approach for medical concept extraction, in: Medical Information Search Workshop (MEDIR) at SIGIR, Pisa, Italy, 2016.

[37] Y. Zha, Y. Yang, R. Li, Z. Hu, AlignScore: Evaluating factual consistency with a unified alignment function, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11328–11348. doi:10.18653/v1/2023.acl-long.634.

[38] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[39] M. Tan, Q. V. Le, EfficientNet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the International Conference on Machine Learning (ICML 2019), 2019, pp. 6105–6114.

[40] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[41] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976. doi:10.1109/CVPR52688.2022.01167.

[42] Q. Jin, W. Kim, Q. Chen, D. C. Comeau, L. Yeganova, W. J. Wilbur, Z. Lu, MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval, Bioinformatics 39 (2023). doi:10.1093/bioinformatics/btad651.

[43] A. L. Beam, B. Kompa, A. Schmaltz, I. Fried, G. Weber, N. Palmer, X. Shi, T. Cai, I. S. Kohane, Clinical concept embeddings learned from massive sources of multimodal medical data, in: Biocomputing 2020, 2019. doi:10.1142/9789811215636_0027.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition

(CVPR 2009), 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[45] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International Conference on Machine Learning, 2022, pp. 12888–12900.

[46] W. Dai, J. Li, D. LI, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, S. Hoi, InstructBLIP: Towards general-purpose vision-language models with instruction tuning, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 49250–49267.

[47] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[48] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, BioBART: Pretraining and evaluation of a biomedical generative language model, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 21st Workshop on Biomedical Language Processing (BioNLP 2022), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 97–109. doi:10.18653/v1/2022.bionlp-1.9.

[49] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, X. Zhai, SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL: https://arxiv.org/abs/2502.14786. arXiv:2502.14786.

[50] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, H. Poon, A multimodal biomedical foundation model trained from fifteen million image–text pairs, NEJM AI 2 (2024). doi:10.1056/AIoa2400640.

[51] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[52] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., LoRA: Low-rank adaptation of large language models, in: Proceedings of the International Conference on Learning Representations (ICLR), volume 1, 2022, p. 3. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[53] ContactDoctor, ContactDoctor-Bio-Medical: A high-performance biomedical language model, https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B, 2024.

[54] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano,

R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[55] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. doi:10.1093/bioinformatics/btz682.

[56] Z. Wang, Z. Wu, D. Agarwal, J. Sun, MedCLIP: Contrastive learning from unpaired medical images and text, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 3876–3887. doi:10.18653/v1/2022.emnlp-main.256.

[57] Meta AI, The llama 4 herd: The beginning of a new era of natively multimodal intelligence, 2025. URL: https://ai.meta.com/blog/llama-4-multimodal-intelligence/, accessed: 2025-06-05.

[58] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report, 2025. URL: https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[59] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu,

M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, Qwen2.5-VL technical report, 2025. URL: https://arxiv.org/abs/2502.13923. arXiv:2502.13923.

[60] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3992–4003. doi:10.1109/ICCV51070.2023.00371.

[61] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8, 2023. URL: https://github.com/ultralytics/ultralytics.

[62] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[63] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[64] A. Binder, G. Montavon, S. Lapuschkin, K. Müller, W. Samek, Layer-wise relevance propagation for neural networks with local renormalization layers, in: Proceedings of the International Conference on Artificial Neural Networks (ICANN 2016), 2016. doi:10.1007/978-3-319-44781-0_8.

A. Full Results

Table 12
Performance of the participating teams in the ImageCLEFmedical 2025 Concept Detection subtask.

Group Name	Run	F1	Secondary F1
AUEB NLP Group	1980	**0.588788**	0.948442
AUEB NLP Group	1981	0.588005	0.950632
AUEB NLP Group	1979	0.587366	0.952266
AUEB NLP Group	1977	0.586759	0.944906
AUEB NLP Group	1982	0.586628	0.950797
AUEB NLP Group	1978	0.586606	0.946598
AUEB NLP Group	1976	0.586443	0.943558
AUEB NLP Group	1975	0.5858	0.938881
AUEB NLP Group	1983	0.585592	0.951541
AUEB NLP Group	1986	0.585395	**0.958925**
AUEB NLP Group	1771	0.584075	0.948869
AUEB NLP Group	1970	0.581914	0.952065
AUEB NLP Group	1973	0.581705	0.946296
AUEB NLP Group	1974	0.580833	0.933403
AUEB NLP Group	1985	0.577306	0.945636
DeepLens	1375	0.576579	0.929936
DeepLens	1725	0.576579	0.929936
DeepLens	1704	0.576411	0.923058
AUEB NLP Group	1984	0.575555	0.944689
DeepLens	1678	0.575385	0.915613
DeepLens	1512	0.575199	0.930402
DeepLens	1677	0.574419	0.922487
DeepLens	1707	0.573931	0.911556
DeepLens	1703	0.572481	0.92337
DeepLens	1513	0.572446	0.930584
DeepLens	1705	0.571973	0.912433
DeepLens	1728	0.571531	0.920061
DeepLens	1514	0.571071	0.916518
DeepLens	1706	0.569956	0.911898
DeepLens	1726	0.569058	0.901976
mapan	1505	0.565985	0.929801
mapan	1658	0.565037	0.925401
mapan	1361	0.564746	0.929101
mapan	1657	0.56174	0.925503
UIT-Oggy	1892	0.561317	0.910382
UIT-Oggy	1894	0.561317	0.910382
DS4DH	1508	0.522459	0.867173
UIT-Oggy	1950	0.5163	0.8268
UIT-Oggy	1971	0.454259	0.719997
UIT-Oggy	1969	0.445994	0.664261
sakthiii	1774	0.400278	0.908151
JJ-VMed	1903	0.398163	0.83292
UMUTeam	1807	0.239768	0.53766
UIT-Oggy	1732	0.197456	0.363732
UMUTeam	1806	0.188454	0.539004
LekshmiscopeVIT	1942	0.149379	0.229757
LekshmiscopeVIT	1928	0.14405	0.221592
LekshmiscopeVIT	1935	0.143856	0.225208
LekshmiscopeVIT	1931	0.14368	0.230574
UIT-Oggy	1387	0.000675	0
UIT-Oggy	1733	0.000408	0.000245

Table 13
Performance of the participating teams in the ImageCLEFmedical 2025 Caption Prediction for the relevance metrics Similarity, BERTScore (Recall), ROUGE-1 and BLEURT.

Group Name	Run	Similarity	BERTScore (Recall)	ROUGE-1	BLEURT	Average
UMUTeam	1681	**0.9270669902**	0.5976714391	**0.2594154991**	0.323028482	**0.5267956026**
UMUTeam	1651	0.927015094	0.5976714391	**0.2594154991**	0.323028482	0.5267826286
DS4DH	1520	0.9016275484	**0.6067308264**	0.2516020917	0.3095523716	0.5173782095
DS4DH	1713	0.901575798	**0.6067308264**	0.2516020917	0.3095523716	0.5173652719
DS4DH	1735	0.8902308449	0.6033844645	0.2420076118	0.3087292771	0.5110880496
DS4DH	1714	0.8900415234	0.6033844645	0.2420076118	0.3087292771	0.5110407192
DS4DH	1946	0.8626234373	0.6010897754	0.2497290254	0.3063632513	0.5049513723
AI Stat Lab	1900	0.8919229234	0.5822740783	0.2439531842	0.3172543037	0.5088511224
AI Stat Lab	1965	0.9008491427	0.5812832767	0.2397171258	0.3185598757	0.5101023552
AI Stat Lab	1951	0.9008150686	0.5812832767	0.2397171258	0.3185598757	0.5100938367
UIT-Oggy	1914	0.8798202189	0.5951423837	0.2535223392	0.3020021864	0.507621782
UIT-Oggy	1922	0.8795908477	0.5951423837	0.2535223392	0.3020021864	0.5075644393
UIT-Oggy	1937	0.8793885703	0.5951423837	0.2535223392	0.3020021864	0.5075138699
UIT-Oggy	1911	0.8790418888	0.5951423837	0.2535223392	0.3020021864	0.5074271995
DS4DH	1662	0.8740206857	0.598339581	0.2339253215	0.3094411571	0.5039316863
AI Stat Lab	1952	0.9011289068	0.5793446009	0.244807675	0.3194920946	0.5111933193
AI Stat Lab	1944	0.851023204	0.5853606745	0.2436787283	0.3178475372	0.499477536
AI Stat Lab	1940	0.8506289832	0.5831456219	0.2422876358	0.3163584231	0.498105166
AI Stat Lab	1947	0.8453687491	0.5801300292	0.2389158623	0.3176919038	0.4955266361
UIT-Oggy	1908	0.8795646934	0.5922596574	0.2474369465	0.2980985607	0.5043399645
UIT-Oggy	1912	0.8795351264	0.5922596574	0.2474369465	0.2980985607	0.5043325727
DS4DH	1902	0.867253701	0.5952384799	0.2295892987	0.3099015828	0.5004957656
UIT-Oggy	1916	0.8794870123	0.5922596574	0.2474369465	0.2980985607	0.5043205442
DS4DH	1525	0.8672293565	0.5952384799	0.2295892987	0.3099015828	0.5004896795
UIT-Oggy	1936	0.8794485511	0.5922596574	0.2474369465	0.2980985607	0.5043109289
UIT-Oggy	1910	0.879448055	0.5922596574	0.2474369465	0.2980985607	0.5043108049
UIT-Oggy	1907	0.8794236755	0.5922596574	0.2474369465	0.2980985607	0.50430471
UIT-Oggy	1906	0.8793601247	0.5922596574	0.2474369465	0.2980985607	0.5042888223
UIT-Oggy	1918	0.8793462955	0.5922596574	0.2474369465	0.2980985607	0.504285365
UIT-Oggy	1905	0.8793214054	0.5922596574	0.2474369465	0.2980985607	0.5042791425
UIT-Oggy	1920	0.8792710703	0.5922596574	0.2474369465	0.2980985607	0.5042665587
UIT-Oggy	1909	0.8792222183	0.5922596574	0.2474369465	0.2980985607	0.5042543457
UIT-Oggy	1913	0.8791593534	0.5922596574	0.2474369465	0.2980985607	0.5042386295
UIT-Oggy	1917	0.8790993466	0.5922596574	0.2474369465	0.2980985607	0.5042236278
UIT-Oggy	1915	0.8789722703	0.5922596574	0.2474369465	0.2980985607	0.5041918587
AI Stat Lab	1941	0.8436483235	0.5823160331	0.2431369998	0.3154446557	0.496136503
AI Stat Lab	1695	0.84912194	0.5774936852	0.2389612648	0.3154152172	0.4952480268
AI Stat Lab	1901	0.8455747573	0.5756069469	0.2351401032	0.3124395254	0.4921903332
DS4DH	1344	0.8324795592	0.5963964174	0.2380531842	0.3054558349	0.4930962489
UIT-Oggy	1224	0.8599589108	0.5886605787	0.2439080163	0.2945956769	0.4967807957
UIT-Oggy	1289	0.8597930593	0.5886605787	0.2439080163	0.2945956769	0.4967393328
UIT-Oggy	1204	0.8597542462	0.5886605787	0.2439080163	0.2945956769	0.4967296295
UIT-Oggy	1219	0.8597205655	0.5886605787	0.2439080163	0.2945956769	0.4967212094
AI Stat Lab	1673	0.8364538401	0.5741285064	0.232773627	0.3129990471	0.4890887552
AI Stat Lab	1729	0.8430352433	0.583511695	0.2331348627	0.3131518531	0.4932084135
AUEB NLP Group	1403	0.7946814175	0.5884477399	0.2176475984	0.302975852	0.4759381519
AI Stat Lab	1948	0.8686241956	0.577509811	0.2173791759	0.31989704	0.4958525556
AUEB NLP Group	1463	0.7942309996	0.5930055156	0.2192303479	0.3013233487	0.4769475529
AUEB NLP Group	1462	0.7940360448	0.593092524	0.2190622343	0.3011381619	0.4768322413
AI Stat Lab	1693	0.8202862609	0.5776584378	0.2269654242	0.3094469583	0.4835892703
AI Stat Lab	1405	0.8222836481	0.5756415949	0.2279223778	0.3094332639	0.4838202212
JJ-VMed	1896	0.8251128696	0.5952554406	0.2388722869	0.3094253234	0.4921664801
JJ-VMed	1953	0.8250990632	0.5952554406	0.2388722869	0.3094253234	0.4921630285

Group Name	Run	Similarity	BERTScore (Recall)	ROUGE-1	BLEURT	Average
AUEB NLP Group	1717	0.7939443954	0.5908580871	0.2174402227	0.3004115043	0.4756635524
AI Stat Lab	1949	0.8605397462	0.5755094226	0.2194102787	0.3179662772	0.4933564312
AUEB NLP Group	1968	0.7885883406	0.5950158604	0.2150092584	0.2925783725	0.472797958
AI Stat Lab	1939	0.815931258	0.5759145452	0.2250088776	0.3089080595	0.4814406851
AI Stat Lab	1759	0.8871127837	0.5714247114	0.2135729257	**0.3267615565**	0.4997179943
AUEB NLP Group	1724	0.7896534529	0.5938921733	0.2121607527	0.2897385987	0.4713612444
AI Stat Lab	1972	0.866789733	0.5741138853	0.2102240107	0.3195298643	0.4926643733
AI Stat Lab	1407	0.8272838955	0.5709968858	0.2220934254	0.3048902393	0.4813161115
DS4DH	1715	0.8517214244	0.5813874618	0.2058265321	0.30820145	0.4867842171
DS4DH	1740	0.8515940058	0.5813874618	0.2058265321	0.30820145	0.4867523624
AI Stat Lab	1769	0.8692591188	0.5751669697	0.2077211375	0.3227973895	0.4937361539
AI Stat Lab	1760	0.874319018	0.5692727362	0.2056533176	0.3228536042	0.493024669
AI Stat Lab	1758	0.8737519664	0.5682426019	0.2050639657	0.3214654774	0.4921310028
AI Stat Lab	1938	0.8627533969	0.5704782898	0.2031508481	0.3197355222	0.4890295143
AI Stat Lab	1757	0.8715797625	0.5687446644	0.2011842945	0.3195840435	0.4902731912
AI Stat Lab	1408	0.7723425438	0.574463241	0.2269994767	0.3093488286	0.4707885225
AUEB NLP Group	1718	0.7844345662	0.5896227542	0.2148468288	0.3043625319	0.4733166703
AI Stat Lab	1943	0.8637755594	0.57226763	0.2015586806	0.3175529241	0.4887886985
AUEB NLP Group	1721	0.7897891572	0.5813948544	0.2138169409	0.311333959	0.4740837278
AUEB NLP Group	1723	0.7917244991	0.577736818	0.217124953	0.312086697	0.4746682417
AUEB NLP Group	1958	0.7725077319	0.5872149247	0.2081816026	0.3010042339	0.4672271233
AUEB NLP Group	1957	0.7721005534	0.5872149247	0.2081816026	0.3010042339	0.4671253286
AUEB NLP Group	1669	0.7406335963	0.5916328686	0.2129158395	0.2991840382	0.4610915857
AUEB NLP Group	1954	0.7348429915	0.5898034716	0.2134433145	0.3028956126	0.4602463476
AUEB NLP Group	1670	0.7361830034	0.5903139834	0.2140566332	0.3065920633	0.4617864208
AUEB NLP Group	1960	0.6777612398	0.545355592	0.1814278054	0.2582794544	0.4157060229
AUEB NLP Group	1722	0.7944573759	0.5441874041	0.1984621327	0.321868093	0.4647437514
AUEB NLP Group	1720	0.7521774647	0.5748713989	0.2050792956	0.3057612504	0.4594723524
AUEB NLP Group	1716	0.6842017212	0.5380596983	0.1594842126	0.2874211724	0.4172917011
AUEB NLP Group	1719	0.7518677714	0.5686053592	0.2018762512	0.3053330915	0.4569206184
AUEB NLP Group	1962	0.653930667	0.5620845325	0.1868267796	0.2585293672	0.4153428366
AUEB NLP Group	1961	0.6648839592	0.5472293767	0.1813993066	0.2637027918	0.4143038585
sakthiii	1890	0.7957256077	0.5552566792	0.1606543695	0.2805677495	0.4480511015
AUEB NLP Group	1963	0.6497840212	0.5599914827	0.1886038364	0.257882338	0.4140654196
AUEB NLP Group	1966	0.6529185566	0.565940854	0.1910083614	0.2770356284	0.4217258501
AUEB NLP Group	1967	0.6326793988	0.5615289665	0.1928296088	0.2850164847	0.4180136147
AUEB NLP Group	1959	0.7378352437	0.5350710091	0.1805768485	0.317518852	0.4427504883
AUEB NLP Group	1402	0.7407298697	0.5469746038	0.1469707576	0.2726336208	0.426827213
UIT-Oggy	1386	0.6518306492	0.3994705669	0.1188318905	0.2903692437	0.3651255876
AI Stat Lab	1245	0.6131234597	0.530113321	0.1799603324	0.2837410142	0.4017345318
CS_Morgan	1815	0.5703600506	0.517985392	0.1598199153	0.238530669	0.3716740067
CS_Morgan	1945	0.4201825029	0.5374972128	0.1360990778	0.2576351634	0.3378534892
CS_Morgan	1817	0.4455569753	0.5295828721	0.08732785488	0.2026097055	0.316269352
CS_Morgan	1955	0.3924821619	0.5089190354	0.1346502289	0.2163842754	0.3131089254
CS_Morgan	1956	0.4506571947	0.5110622933	0.0919092927	0.20395449	0.3143958177

Table 15
Performance of the participating teams in the ImageCLEFmedical 2025 Caption Prediction for the factuality metrics UMLS Concept F1-scoreand AlignScore.

Group Name	Run	UMLS Concept F1	AlignScore	Average
UMUTeam	1681	**0.181597074**	0.1375067978	**0.1595519359**
UMUTeam	1651	**0.181597074**	0.1375067978	**0.1595519359**
DS4DH	1520	0.1681505412	0.1417267149	0.154938628
DS4DH	1713	0.1681505412	0.1417267149	0.154938628

Group Name	Run	UMLS Concept F1	AlignScore	Average
DS4DH	1735	0.1607845314	0.1300015817	0.1453930566
DS4DH	1714	0.1607845314	0.1300015817	0.1453930566
DS4DH	1946	0.1545091995	0.12863346	0.1415713298
AI Stat Lab	1900	0.152401917	0.1213079496	0.1368549333
AI Stat Lab	1965	0.1485685306	0.1161946781	0.1323816043
AI Stat Lab	1951	0.1485685306	0.1161946781	0.1323816043
UIT-Oggy	1914	0.1672053088	0.1020638557	0.1346345822
UIT-Oggy	1922	0.1672053088	0.1020638557	0.1346345822
UIT-Oggy	1937	0.1672013955	0.1020638557	0.1346326256
UIT-Oggy	1911	0.1672053088	0.1020638557	0.1346345822
DS4DH	1662	0.1554253529	0.1175877731	0.136506563
AI Stat Lab	1952	0.1478996923	0.107244806	0.1275722492
AI Stat Lab	1944	0.1535761911	0.1233331983	0.1384546947
AI Stat Lab	1940	0.1539626853	0.121640157	0.1378014212
AI Stat Lab	1947	0.1520396375	0.1264092892	0.1392244633
UIT-Oggy	1908	0.1671403074	0.08817711816	0.1276587128
UIT-Oggy	1912	0.1671403074	0.08817711816	0.1276587128
DS4DH	1902	0.1519294103	0.1110485002	0.1314889553
UIT-Oggy	1916	0.1671403074	0.08817711816	0.1276587128
DS4DH	1525	0.1519294103	0.1110485002	0.1314889553
UIT-Oggy	1936	0.1671403074	0.08817711816	0.1276587128
UIT-Oggy	1910	0.1671403074	0.08817711816	0.1276587128
UIT-Oggy	1907	0.1671442207	0.08817711816	0.1276606694
UIT-Oggy	1906	0.1671403074	0.08817711816	0.1276587128
UIT-Oggy	1918	0.1671442207	0.08817711816	0.1276606694
UIT-Oggy	1905	0.1671403074	0.08817711816	0.1276587128
UIT-Oggy	1920	0.1671403074	0.08817711816	0.1276587128
UIT-Oggy	1909	0.1671442207	0.08817711816	0.1276606694
UIT-Oggy	1913	0.1671442207	0.08817711816	0.1276606694
UIT-Oggy	1917	0.1671442207	0.08817711816	0.1276606694
UIT-Oggy	1915	0.1671403074	0.08817711816	0.1276587128
AI Stat Lab	1941	0.1512932382	0.1149847237	0.1331389809
AI Stat Lab	1695	0.1486805908	0.1167159104	0.1326982506
AI Stat Lab	1901	0.1479159041	0.1145724355	0.1312441698
DS4DH	1344	0.1409897282	0.1191117719	0.1300507501
UIT-Oggy	1224	0.1591619369	0.0917596987	0.1254608178
UIT-Oggy	1289	0.1591619369	0.0917596987	0.1254608178
UIT-Oggy	1204	0.1591619369	0.0917596987	0.1254608178
UIT-Oggy	1219	0.1591580237	0.0917596987	0.1254588612
AI Stat Lab	1673	0.1439318728	0.1219713989	0.1329516358
AI Stat Lab	1729	0.1391716321	0.1075658474	0.1233687398
AUEB NLP Group	1403	0.1428841514	0.132528786	0.1377064687
AI Stat Lab	1948	0.1339785951	0.09510961553	0.1145441053
AUEB NLP Group	1463	0.1418998018	0.1229888817	0.1324443418
AUEB NLP Group	1462	0.1418848072	0.1230587846	0.1324717959
AI Stat Lab	1693	0.1403268954	0.1101189418	0.1252229186
AI Stat Lab	1405	0.138163306	0.1117387904	0.1249510482
JJ-VMed	1896	0.1365865963	0.09636041502	0.1164735057
JJ-VMed	1953	0.1365896034	0.09636041502	0.1164750092
AUEB NLP Group	1717	0.1428581972	0.1212715175	0.1320648574
AI Stat Lab	1949	0.135009399	0.09302195493	0.114015677
AUEB NLP Group	1968	0.1416362042	0.1249606724	0.1332984383
AI Stat Lab	1939	0.1384932591	0.1104100202	0.1244516397
AI Stat Lab	1759	0.1339573047	0.07731010822	0.1056337064
AUEB NLP Group	1724	0.1420931805	0.1256619378	0.1338775591
AI Stat Lab	1972	0.1298322027	0.08978797749	0.1098100901
AI Stat Lab	1407	0.1319701123	0.1008284354	0.1163992738

Group Name	Run	UMLS Concept F1	AlignScore	Average
DS4DH	1715	0.1317131746	0.09011638678	0.1109147807
DS4DH	1740	0.1317131746	0.09011638678	0.1109147807
AI Stat Lab	1769	0.1281333934	0.07762737151	0.1028803824
AI Stat Lab	1760	0.1285401193	0.07700460129	0.1027723603
AI Stat Lab	1758	0.1275018379	0.07555846518	0.1015301515
AI Stat Lab	1938	0.1236428821	0.08431520399	0.103979043
AI Stat Lab	1757	0.125356389	0.07871790992	0.1020371495
AI Stat Lab	1408	0.120136846	0.1221238774	0.1211303617
AUEB NLP Group	1718	0.1332347603	0.1031040788	0.1181694195
AI Stat Lab	1943	0.1242196908	0.0795414926	0.1018805917
AUEB NLP Group	1721	0.1264694028	0.1061366883	0.1163030455
AUEB NLP Group	1723	0.1318383455	0.09723908109	0.1145387133
AUEB NLP Group	1958	0.1313131377	0.1048361156	0.1180746266
AUEB NLP Group	1957	0.1313131377	0.1048361156	0.1180746266
AUEB NLP Group	1669	0.1281878816	0.1071699227	0.1176789021
AUEB NLP Group	1954	0.1269069081	0.1023208849	0.1146138965
AUEB NLP Group	1670	0.126466453	0.09754210518	0.1120042791
AUEB NLP Group	1960	0.1038346088	**0.20580824**	0.1548214244
AUEB NLP Group	1722	0.1210673216	0.08515363201	0.1031104768
AUEB NLP Group	1720	0.1149024835	0.09374551904	0.1043240013
AUEB NLP Group	1716	0.08085937435	0.2013026187	0.1410809965
AUEB NLP Group	1719	0.1129864203	0.08519079867	0.09908860949
AUEB NLP Group	1962	0.103754184	0.1684078371	0.1360810105
AUEB NLP Group	1961	0.09976216893	0.1706447497	0.1352034593
sakthiii	1890	0.1094381488	0.09281623394	0.1011271914
AUEB NLP Group	1963	0.1021655207	0.1626846656	0.1324250931
AUEB NLP Group	1966	0.09722323349	0.1471662273	0.1221947304
AUEB NLP Group	1967	0.09908635031	0.1305662774	0.1148263139
AUEB NLP Group	1959	0.09507913073	0.07194208767	0.0835106092
AUEB NLP Group	1402	0.09692104503	0.09017681921	0.09354893212
UIT-Oggy	1386	0.09659835594	0.1357863288	0.1161923424
AI Stat Lab	1245	0.06223023044	0.09672321705	0.07947672375
CS_Morgan	1815	0.0740537927	0.1086801355	0.09136696409
CS_Morgan	1945	0.02326776867	0.07248510048	0.04787643457
CS_Morgan	1817	0.0244483798	0.09559641568	0.06002239774
CS_Morgan	1955	0.01707103176	0.08739293675	0.05223198425
CS_Morgan	1956	0.02251056628	0.07620089553	0.04935573091