# Overview of the MEDIQA-CORE 2026 Task 2 on Radiology Report Discrepancy Assessment

Asma Ben Abacha[1,*,†], Zhaoyi Sun[2,†], Wen-wai Yim[1], Fei Xia[2] and Meliha Yetisgen[2]

[1]Microsoft, Health and Life Sciences AI, Redmond, WA, USA
[2]University of Washington, Seattle, WA, USA

**Abstract**

Radiology report discrepancy assessment is an important clinical task aimed at identifying and characterizing differences between preliminary and revised radiology reports. Such discrepancies may range from minor clarifications to clinically significant errors that directly affect patient management. To advance research in this area, we organized MEDIQA-CORE 2026 Task 2, a new shared task on multimodal radiology report review and discrepancy analysis based on the RADAR benchmark. The task required participating systems to jointly reason over 3D abdominal CT volumes, preliminary radiology reports, and candidate report edits in order to predict agreement with imaging findings, clinical severity, and edit type. The shared task attracted a diverse range of approaches, including prompt-based large language model pipelines, multimodal transformer architectures, and hybrid vision-language systems. Results demonstrate that carefully designed prompting strategies can compete with substantially more complex image-grounded architectures, while multimodal methods continue to provide complementary clinical context for discrepancy understanding. In this paper, we describe the task setup, dataset, evaluation framework, participating systems, and official results, and discuss emerging challenges and future directions for clinically grounded multimodal radiology discrepancy assessment.

**Keywords**

Radiology Report Review, Multimodal Medical AI, Clinical Discrepancy Assessment

## 1. Introduction

Radiology discrepancy assessment is clinically important because report inaccuracies can directly affect diagnosis, treatment planning, and patient management. Proposed report edits may range from simple clarifications to corrections of clinically significant findings, requiring systems to reason not only about semantic consistency between reports and edits, but also about whether the modification is supported by the underlying medical image and how clinically important the discrepancy may be. This makes the task substantially more challenging than conventional report generation or text classification settings.

Recent advances in large language models and multimodal vision-language systems have led to substantial progress in medical image understanding and clinical report generation. However, comparatively less attention has been devoted to radiology report review and discrepancy analysis, where systems must evaluate whether a proposed report modification is factually supported, clinically meaningful, and semantically appropriate. Such tasks require fine-grained multimodal reasoning over both volumetric imaging data and structured clinical language.

In the context of the MEDIQA series of shared tasks[1] in medical AI and the ImageCLEF 2026 lab [1], we organized MEDIQA-CORE 2026 Task 2 on multimodal radiology report review and discrepancy assessment. To support research in this area, the task uses the new RADAR dataset [2], a new multimodal benchmark for radiology report discrepancy analysis. The dataset consists of expert-annotated abdominal CT studies curated to reflect authentic trainee-to-attending report revision workflows. Each

*CLEF 2026 Working Notes, September 21–24, 2026, Jena, Germany*
*Corresponding author.
†These authors contributed equally.
abenabacha@microsoft.com (A. Ben Abacha); zhaoyis@uw.edu (Z. Sun); yimwenwai@microsoft.com (W. Yim); fxia@uw.edu (F. Xia); melihay@uw.edu (M. Yetisgen)
0000-0001-6312-9387 (A. Ben Abacha); 0009-0003-8197-1465 (Z. Sun); 0000-0001-9011-0817 (W. Yim); 0000-0001-6015-6064 (F. Xia); 0000-0001-9919-9811 (M. Yetisgen)


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

sample includes a volumetric abdominal CT examination, a preliminary radiology report, and a candidate report edit. Participating systems were required to jointly predict (1) agreement between the edit and the imaging findings, (2) the clinical severity of the discrepancy, and (3) the edit type.

The shared task attracted a diverse range of approaches, including prompt-based large language model systems, multimodal transformer architectures, and hybrid vision-language frameworks combining clinical text encoders with volumetric CT representations. The results demonstrate the growing capabilities of multimodal and generative AI systems for clinically grounded reasoning, while also highlighting persistent challenges in severity calibration and fine-grained discrepancy understanding.

In this overview paper, we describe the task, dataset, baseline system, evaluation setup, participating systems, and official results of MEDIQA-CORE 2026 Task 2, and discuss emerging directions for multimodal radiology discrepancy analysis.

## 2. Task Description

MEDIQA-CORE 2026 Task 2[2] focuses on multimodal radiology report review and discrepancy assessment. The task is designed to model a realistic clinical workflow in which preliminary radiology reports drafted by junior radiologists are subsequently reviewed and revised by attending radiologists. These revisions range from minor clarifications to clinically significant corrections and provide a realistic source of report discrepancies that can be used to evaluate image-grounded reasoning. Given a 3D abdominal CT examination, a preliminary radiology report, and a candidate report edit, participating systems must determine whether the proposed modification is clinically appropriate and assess the nature of the discrepancy. Such capabilities are directly relevant to AI-assisted radiology report verification, quality assurance, discrepancy analysis, and trainee education.

Each sample in the dataset consists of three primary components: (1) a volumetric abdominal CT scan, (2) a preliminary radiology report describing the study findings, and (3) a candidate edit representing a proposed modification to the report text. Systems are required to jointly reason over imaging and textual information in order to evaluate the edit across three subtasks:

1. **Agreement Classification:** determine whether the proposed edit is supported by the CT findings.
2. **Severity Assessment:** estimate the clinical significance of the discrepancy introduced or corrected by the edit.
3. **Edit Type Classification:** categorize the edit as a correction, addition, or clarification.

The task emphasizes fine-grained clinical reasoning and multimodal alignment between volumetric imaging data and radiology report text. Successful systems must distinguish between edits that improve factual correctness, edits that introduce clinically meaningful errors, and edits that only refine or clarify existing interpretations. This requires not only semantic understanding of clinical language, but also grounding textual claims in the underlying CT examination.

By framing radiology review as a multimodal reasoning problem, the task aims to encourage the development of clinically grounded AI systems capable of assisting radiologists during report verification and quality assurance.

## 3. Data Description

We used the RADAR benchmark [2], a new multimodal dataset for radiology report discrepancy analysis[3]. RADAR consists of expert annotated abdominal CT studies curated to reflect authentic trainee-to-attending report revision workflows. The dataset includes 3D volumetric abdominal CT examinations, preliminary radiology reports, and candidate edits representing proposed modifications

[2] https://www.imageclef.org/2026/medical/mediqa-core

[3] Participants were granted access to the competition dataset after completing the ImageCLEF registration form and the UW Data Usage Agreement, which required approval by the University of Washington. More information about requesting access to the dataset after the competition can be found at the following link: https://github.com/GeraldSun/RADAR

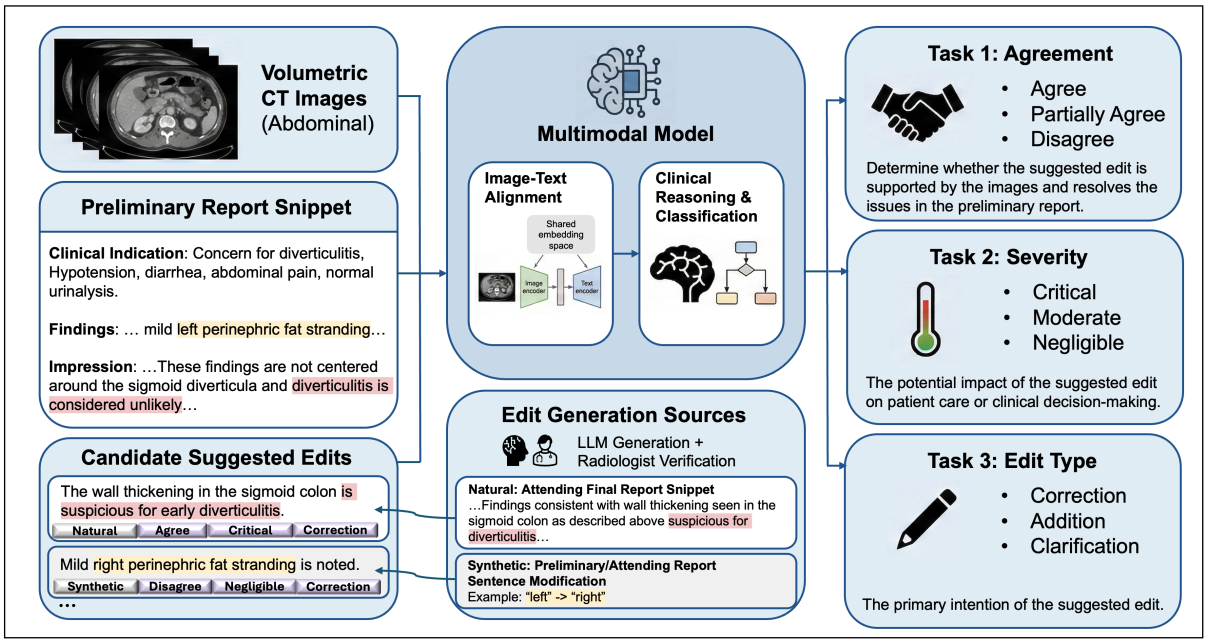


**Figure 1:** RADAR workflow illustrating the benchmark pipeline and its constituent tasks.

to the original report text. Figure 1 presents the RADAR workflow and summarizes the tasks considered in this benchmark.

The benchmark covers 50 abdominal CT examinations (20 validation, 30 test), balanced across daytime and overnight ED cases, and comprising 319 3D CT image series (volumes) and 45,838 slices in total. The test set contains 153 natural candidate edits and 274 mixed edits, where the "Mixed" set augments the natural revisions with expert-verified synthetic "Disagree" edits introduced to balance the agreement-label distribution and to assess model robustness to unsupported edits.

## 4. Baseline Methods

The baseline methods[4] evaluate several multimodal foundation models for detecting and reasoning about discrepancies between CT images and radiology report edits. The baselines include Google's Gemini-2.5-Pro and Gemini-3-Pro, as well as Alibaba's Qwen3.5-plus, tested under different visual input settings such as 10, 20, or 50 sampled image slices and full-video representations of scans. The framework also includes a series-selection step that identifies the most relevant CT series for each edit before inference, enabling standardized comparison of agreement detection, severity prediction, and edit-type classification performance across models [2]. The final baseline model presented here is the Gemini-3-Pro with the 50 slice setting.

## 5. Evaluation Metrics

Each candidate report edit was evaluated along three dimensions: agreement, clinical severity, and edit type. These dimensions were designed to capture both the factual validity of the proposed modification and its potential clinical importance within the radiology report review workflow.

**Agreement Classification.** The agreement metric measures whether a proposed edit is consistent with the reviewing physician's assessment using three categories:

[4]Code: https://github.com/GeraldSun/RADAR

- **Agree:** the edit is fully accurate and introduces a clinically meaningful correction or missing finding.
- **Partially Agree:** the edit is generally correct but may be redundant, overly nitpicky, imprecise, or clinically minor.
- **Disagree:** the edit contains factual errors, misinterprets the imaging findings, or is unnecessary.

**Severity Classification.** The severity metric evaluates the potential clinical impact associated with including or omitting the edit:

- **Critical:** the discrepancy could lead to immediate life-threatening harm or dangerous unnecessary intervention.
- **Moderate:** the discrepancy could delay non-emergent care or result in unnecessary follow-up.
- **Negligible:** the discrepancy would have little or no practical effect on patient management.

**Edit Type Classification.** The edit-type metric categorizes each edit according to its intended role in revising the report:

- **Correction:** the edit fixes an incorrect finding in the preliminary report.
- **Addition:** the edit adds a previously omitted finding.
- **Clarification:** the edit refines or clarifies an imprecise statement already present in the report.

The official evaluation was conducted using the AI4MediaBench[5] platform and reported performance on two dataset splits. The **Mixed** split contained all examples, including both natural and synthetic edits, while the **Natural** split consisted only of naturally occurring report edits. The natural-edit subset was hidden during the test phase, and participants were not provided with information identifying which test examples belonged to this subset.

For each split, performance was assessed using four metrics: agreement accuracy, severity accuracy, edit-type accuracy, and a composite score combining the three subtasks. The composite score was defined as a joint metric requiring a fuzzy match on agreement and exact matches on severity and edit type. For the agreement component, "Agree" and "Partially Agree" were treated as equivalent, while Disagree remained distinct. For severity and edit type, predictions were required to exactly match the reference labels. This resulted in a total of eight reported metrics:

- Mixed Agreement Accuracy
- Mixed Severity Accuracy
- Mixed Edit-Type Accuracy
- Mixed Composite Score
- Natural Agreement Accuracy
- Natural Severity Accuracy
- Natural Edit-Type Accuracy
- Natural Composite Score

The primary ranking metric for the shared task was the Mixed Composite Score.

## 6. Participating Teams

MEDIQA-CORE 2026 Task 2 attracted 22 registered teams from different countries. Among them, four finalist teams submitted their runs following the challenge rules. Table 1 presents the teams that participated in the challenge. We limited the number of submitted runs to 10 runs per team.

[5]Leaderboard: https://ai4media-bench.aimultimedialab.ro/competitions/7/

**Table 1**
Participating teams, papers, and codes.

| | Team | Username | Affiliation | Paper | Code |
|---|---|---|---|---|---|
| 1 | Brihadaranyaka | anubhav27 | Independent Researcher, India | [3] | [1] |
| 2 | MediFact | nadiasaeed | Quaid-i-Azam University, Islamabad, Pakistan | [4] | [2] |
| 3 | ClinicalVLM-TW | tzuyi_li | National Chung Hsing University, Taichung & National Taiwan University Hospital, Taipei, Taiwan | [5] | [3] |
| 4 | IReL_IIT(BHU) | krishnatewari | Indian Institute of Technology & Banasthali Vidyapith, India | [6] | [4] |

[1] https://github.com/AnubhavBharadwaaj/mediqatask1task2analysisnotebbok
[2] https://github.com/NadiaSaeed/Medifact-MEDIQA-CORE-Task-2-2026
[3] https://github.com/Alice-LTY/mediqa-core-task2
[4] https://github.com/cse-iitbhu/Report-Discrepancy-Summarization

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

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

**Table 3**
MEDIQA-CORE 2026 Task 2 leaderboard results on the Natural Set.

| Rank | Team | ID | Composite (Natural) | Agreement (Natural) | Severity (Natural) | Edit Type (Natural) |
|---|---|---|---|---|---|---|
| 1 | Brihadaranyaka | 1028 | **0.58** | **0.86** | **0.76** | **0.72** |
| 1 | MediFact | 1043 | **0.58** | 0.76 | **0.76** | **0.72** |
| – | *MEDIQA Organizers* | 797 | 0.48 | 0.63 | 0.61 | 0.84 |
| 3 | ClinicalVLM-TW | 1037 | 0.29 | 0.49 | 0.46 | 0.65 |
| 4 | IReL_IIT(BHU) | 960 | 0.08 | 0.24 | **0.76** | 0.14 |

## 7. Official Results

The main results are summarized in Table 2 and Table 3. Our baseline systems were not included in the official ranking.

On the Mixed dataset (Table 2), our baseline system achieved the strongest overall performance with a composite score of 0.39. Among the participating teams, Brihadaranyaka ranked first with a composite score of 0.33, followed closely by MediFact at 0.32. The top-performing submissions achieved relatively balanced performance across agreement, severity, and edit-type prediction, while lower-ranked systems showed substantial degradation in one or more subtasks.

On the Natural dataset (Table 3), two teams (Brihadaranyaka and MediFact) tied for first place with a composite score of 0.58, outperforming our baseline system (0.48). Compared with the Mixed dataset, most systems achieved higher agreement and severity scores on the Natural set, suggesting that naturally occurring report discrepancies may be easier to model than the broader mixed distribution.

## 8. System Descriptions

**Brihadaranyaka.** The Brihadaranyaka team proposed a single-call zero-shot pipeline for the three classification tasks (agreement, severity, and edit type). Their system used `claude-sonnet-4-6` at temperature 0 with strictly text-only input consisting of the preliminary report and candidate edit, deliberately excluding access to the CT volume. The core contribution was a carefully engineered system prompt incorporating an explicit severity-decoupling rule (an incorrect edit may still be clinically negligible), a counterfactual care-impact heuristic ("would the patient's next 24-48 hours of care change if the edit were added or removed?"), and several guardrails designed to prevent overuse of the moderate severity class. The team submitted four prompt variants, differing only in the system prompt. Their best-performing runs achieved 1st place on the Mixed Set leaderboard and tied for 1st on the Natural Set leaderboard among finalist teams, outperforming the baseline system. Their paper further reports statistical comparisons between prompt variants, analyzes the dominant moderate-class overprediction failure mode, and presents additional post-competition experiments with different models and ensemble-style prompting strategies [3].

**MediFact.** The MediFact team proposed an interaction-aware multimodal framework for radiology report discrepancy analysis that jointly leveraged abdominal CT volumes, preliminary reports, and candidate edits. The proposed system achieved second place on the Mixed Set and first place on the Natural Set. Their approach combined textual and imaging representations: clinical text embeddings were extracted using the `DeBERTa-v3` transformer with mean-pooled contextual representations, while CT imaging features were obtained using a lightweight 3D CNN. To support fine-grained discrepancy reasoning, the framework explicitly modeled semantic relationships between the preliminary report and candidate edit through difference vectors and element-wise interaction features prior to multimodal fusion. The architecture further incorporated CT preprocessing, modality-balanced feature fusion, layer normalization, and weighted multitask optimization across the agreement, severity, and edit-type prediction objectives [4].

**ClinicalVLM-TW**. The ClinicalVLM-TW team developed an Evidence-Anchored, Confidence-Gated, cross-backbone multi-agent VLM pipeline for radiology report discrepancy assessment that emulates Radiology Resident Reasoning. The pipeline decomposes Radiology Resident Reasoning into three per-edit sub-processes: (i) Multi-Slice Reading via key-slice selection across the 3D CT volume, (ii) Rule-Augmented Clinical Reasoning via in-context conditioning on thirteen procedural rules, and (iii) retrieval-augmented evidence injection via OpenEvidence per-edit lookup. The Evidence-Anchored property is realized by sub-process (iii), while the Confidence-Gated property is realized through Per-Field Hybrid Prompt Selection (per-label argmax over single-prompt variants evaluated on the development split) layered with a Multi-Agent Ensemble over four LLM backbones (Claude Opus, Claude Sonnet, GPT-4o, Gemini 2.5), Multi-Source Severity Voting, and a Rule-Based Refinement Mechanism. Ablation studies cover five Multi-Slice Reading strategies and three Multi-Agent Ensemble strategies, each accompanied by root-cause analysis. The proposed system attained a development-set composite score of 0.549 and a Mixed Set test composite score of 0.22, with the agreement metric (0.53) ranking highest among finalist teams. In addition, an inter-rater development-split sub-study involving four physician partners from four Taiwanese teaching hospitals yielded a severity kappa of 0.27 and produced two physician-derived community resources: 29 per-case clinical analyses and a consolidated set of consensus and specialty-divergent severity anchor scenarios. In addition, a blinded calibration sub-study on a development subset observed fair inter-physician agreement on severity (Cohen's $\kappa \approx 0.34$). The team also released the inference pipeline, field-optimized prompt configurations, prediction artifacts, and four curated research resources: per-edit OpenEvidence query logs, clinical-rationale annotations, per-case clinical analyses, and consensus severity anchor scenarios. [5].

**IReL_IIT(BHU).** The IReL_IIT(BHU) team built a hybrid multimodal framework for radiology report discrepancy analysis in MEDIQA-CORE Task 2. Their approach jointly addressed agreement classification, severity prediction, and edit-type classification using transformer-based textual representations combined with multimodal feature fusion. Clinical indications, radiology reports, and candidate edits were encoded using `PubMedBERT`, while abdominal CT scan volumes were processed with a 3D CNN

to obtain complementary imaging representations. The framework adopted a hierarchical classification strategy in which threshold-based decision refinement was applied across the agreement, severity, and edit-type subtasks. To improve robustness and mitigate class imbalance, the team additionally introduced synthetic radiology edit augmentation during training. Experimental results demonstrated the effectiveness of combining transformer-based clinical language modeling with multimodal imaging features for radiology report review and discrepancy understanding [6].

## 9. Conclusion

MEDIQA-CORE 2026 Task 2 demonstrated the effectiveness of both prompt-based large language model systems and multimodal vision-language architectures for radiology report discrepancy assessment. The finalist teams explored diverse strategies, including zero-shot prompting without image access, interaction-aware multimodal fusion, hybrid transformer-CNN frameworks combining clinical text and CT imaging features, and evidence-anchored multi-agent VLM pipelines incorporating multi-slice reading, rule-guided reasoning, and retrieval-augmented evidence grounding. Across systems, accurate severity calibration and robust discrepancy reasoning emerged as central challenges. The results suggest that carefully designed prompting approaches can compete with substantially more complex image-grounded models, while multimodal methods continue to provide complementary clinical context for automated discrepancy detection. Future work may benefit from combining the strengths of both paradigms: the interpretability and reasoning flexibility of large language models with the grounding and contextual detail provided by medical imaging, as well as structured reasoning and evidence-aware ensemble strategies for more reliable discrepancy assessment.

## Acknowledgements

We would like to thank Paul Vozila for his feedback, the ImageCLEF organizers for their support, our annotators for their efforts, and the participating teams for their strong engagement and interesting approaches, which contributed to the success of the shared task.

## Declaration on Generative AI

During the preparation of this work, ChatGPT was used to assist with proofreading in some sections of the manuscript. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References

[1] B. Ionescu, H. Müller, D. Stanciu, A. Radu, R. Bolborici, M. Negru, A. Ene, V. Vasilescu, A.-A. Nicolae, L. Ştefan, M. Constantin, M. Dogariu, A. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, R. Brüngel, L. Reinartz, H. Schäfer, C. S. Schmidt, B. Bracke, P. Nath, B. Eryılmaz, M. Hjuler, D. Fabre, C. Lemaire, B. Lecouteux, D. Schwab, D. Dimitrov, M. S. Hee, M. Ahsan, S. Ahmad, D. Zlatkova, G. Pachov, Z. Xie, P. Nakov, I. Koychev, J. E. Heras Rivera, D. K. Low, W. Yim, J. Ruzevick, D. Child, M. Kurt, Z. Sun, F. Xia, M. Yetisgen, A. Radzhabov, Y. Prokopchuk, V. Kovalev, D. Karpenka, S. A. Hicks, S. Gautam, M. A. Riegler, V. Thambawita, P. Halvorsen, M. El Sakka, J. Mothe, A. Băicoianu, C. Florea, M. Ivanovici, Overview of imageclef 2026: Multimodal challenges in medicine, science, agritech, and security, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, 2026.

[2] Z. Sun, M. Jagtiani, W.-w. Yim, F. Xia, M. Gunn, M. Yetisgen, A. Ben Abacha, Radar: A multimodal benchmark for 3d image-based radiology report review, in: Medical Image Computing and Computer Assisted Intervention - MICCAI 2026 - 29th International Conference, Strasbourg, France, 2026. URL: https://doi.org/10.48550/arXiv.2603.06681.
[3] A. Kumar, Brihadaranyaka at mediqa-core-task-2 2026: A text-only zero-shot pipeline with explicit severity-decoupling rules for radiology report discrepancy assessment, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[4] N. Saeed, Medifact at mediqa-core-task-2 2026: Interaction-aware multimodal fusion for fine-grained radiology report discrepancy analysis, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[5] T.-Y. Li, H.-L. Lin, Y.-C. Tu, P. Y. Chan, T.-Y. Hsieh, S.-J. Wang, Clinicalvlm-tw at mediqa-core-task-2 2026: Emulating radiology resident reasoning via per-field hybrid prompt selection in an evidence-anchored, confidence-gated vlm pipeline for radiology report discrepancy assessment, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[6] K. Tewari, S. Verma, S. Pal, Irel_iit(bhu) at mediqa-core-task-2 2026: Hybrid multimodal learning for radiology report discrepancy analysis, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.