# Overview of the MEDIQA-CORE 2026 Task 1 on Brain Tumor Subtype Classification

Asma Ben Abacha[1,*,†], Juampablo E. Heras Rivera[2,†], Daniel K. Low[2], Wen-wai Yim[1], Jacob Ruzevick[2], Daniel D. Child[2] and Mehmet Kurt[2]

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

**Abstract**

This paper presents an overview of MEDIQA-CORE 2026 Task 1 on multimodal brain tumor subtype classification under incomplete-modality settings, organized as part of the MEDIQA shared tasks and the ImageCLEF 2026 lab. The task is based on the CoRe-BT benchmark, which combines multi-sequence brain MRI volumes, histopathology whole-slide imaging, and radiology reports collected from multiple de-identified clinical cohorts. To reflect realistic diagnostic workflows, the benchmark includes both fully multimodal cases and cases with missing modalities, requiring systems to remain robust when histopathology data are unavailable. Participants addressed three classification objectives: molecular grouping, WHO grade prediction, and low-grade versus high-grade glioma classification. A total of five finalist teams submitted systems exploring diverse multimodal approaches, including fusion architectures, embedding-based learning, classical machine-learning ensembles, and large language model-guided reasoning and refinement. Results demonstrate the effectiveness of integrating complementary radiological, pathological, and textual information for brain tumor characterization, while also highlighting the continued challenges associated with fine-grained WHO grade prediction under heterogeneous and incomplete multimodal inputs. The shared task establishes a benchmark for studying multimodal medical reasoning in realistic clinical settings and provides a foundation for future research on robust and clinically deployable brain tumor classification systems.

**Keywords**

Multimodal Learning, Brain Tumor Classification, Medical Vision-Language Models

## 1. Introduction

Accurate brain tumor classification is central to clinical decision-making, guiding treatment selection and prognosis estimation. In clinical workflows, diagnosis is informed by complementary sources of information, including magnetic resonance imaging (MRI), histopathological examination, and associated reports. MRI is typically available early and provides non-invasive characterization of tumor morphology and tissue properties across multiple sequences (e.g., T1, T1c, T2, FLAIR). In contrast, histopathology offers definitive cellular-level insights but is invasive and often delayed due to tissue acquisition and processing. This temporal and modality imbalance motivates the development of computational methods that can leverage available data at different stages of the diagnostic workflow.

Recent work in multimodal learning for medical applications has increasingly focused on integrating heterogeneous data sources to improve predictive performance and robustness. Large-scale imaging benchmarks such as the BraTS challenges have driven substantial progress in MRI-based brain tumor analysis, including segmentation and downstream clinical tasks such as survival prediction [1, 2, 3]. Beyond imaging alone, cross-modal fusion of heterogeneous biomedical data has shown promise for tasks such as outcome prediction [4] and tumor characterization [5]. Complementary efforts such as

*CLEF 2026 Working Notes, September 21–24, 2026, Jena, Germany*
*Corresponding author.
†These authors contributed equally.
abenabacha@microsoft.com (A. Ben Abacha); jehr@uw.edu (J. E. Heras Rivera); dalow@uw.edu (D. K. Low); yimwenwai@microsoft.com (W. Yim); ruzevick@uw.edu (J. Ruzevick); dchild@uw.edu (D. D. Child); mkurt@uw.edu (M. Kurt)
0000-0001-6312-9387 (A. Ben Abacha); 0000-0002-0205-6329 (J. E. Heras Rivera); 0000-0001-9011-0817 (W. Yim); 0000-0002-4309-0144 (D. D. Child); 0000-0002-4309-0144 (M. Kurt)

the Computational Precision Medicine Radiology-Pathology challenge further demonstrate the value of jointly leveraging imaging and histopathology for brain tumor classification [6]. More recently, medical vision-language models have explored aligning imaging data with associated textual reports [7, 8]. Despite these advances, most existing methods assume complete, synchronized, and fully available multimodal inputs during both training and inference. This assumption limits their applicability in real-world clinical settings, where modalities are frequently inaccessible, missing, incomplete, or acquired asynchronously [9].

A key challenge in multimodal tumor classification lies in the intrinsic heterogeneity of the data. MRI captures macroscopic anatomical and functional information, histopathology provides high-resolution microscopic detail, and reports encode expert knowledge in unstructured text. These modalities differ not only in scale and representation but also in their clinical roles, requiring models that can effectively reason across domains while remaining robust to incomplete inputs. Furthermore, relatively few existing benchmarks explicitly evaluate robustness to missing modalities or temporally staggered clinical data availability.

To address these challenges, we introduce MEDIQA-CORE Task 1 as part of the MEDIQA shared tasks[1] and the ImageCLEF 2026 lab [10]. The task focuses on tumor type prediction from multimodal clinical data and is designed to reflect realistic diagnostic conditions, including variability in modality availability at inference time. Specifically, the task evaluates systems that integrate heterogeneous inputs while remaining robust to incomplete multimodal evidence, thereby better reflecting realistic clinical decision-making scenarios. The official evaluation was conducted using the AI4MediaBench[2] platform.

## 2. Task Description

MEDIQA-CORE 2026[3] Task 1 focuses on multimodal brain tumor classification using four MRI sequences (T1, T1c, T2, and FLAIR), H&E-stained histopathology whole-slide images, and radiology reports.

Participants are required to design models capable of integrating radiological, pathological, and textual information to predict tumor subtype and grade under realistic incomplete-modality settings. To reflect clinical workflows in which histopathological confirmation may be delayed or unavailable, the dataset includes both fully multimodal cases and cases with missing modalities containing only MRI and radiology reports. Modality availability is explicitly specified for each sample.

The dataset contains the following input configurations:

- MRI + histopathology images + radiology reports
- MRI images + radiology reports only

The challenge comprises three classification objectives:

- **Task 1: Molecular Grouping** (`level1_label`) predicts coarse biologically meaningful molecular subtypes:
  - **0 – GBM / IDH-wt:** Glioblastoma and IDH-wildtype high-grade glioma
  - **1 – IDH-mutant Diffuse:** Astrocytoma (IDH-mutant) and oligodendroglioma
  - **2 – Pediatric-type:** Midline glioma (H3-altered), pilocytic, and MAPK-driven tumors
- **Task 2: LGG vs. HGG** (`lgghgg_label`) distinguishes low-grade glioma from high-grade glioma:
  - **0 – LGG:** Low-grade glioma (WHO Grade 2)
  - **1 – HGG:** High-grade glioma (WHO Grades 3–4)
- **Task 3: Histological Grade** (`who_grade_label`) predicts WHO tumor grade:
  - **0 – WHO Grades 1–2**
  - **1 – WHO Grade 3**
  - **2 – WHO Grade 4**

[1] https://sites.google.com/view/mediqa-shared-tasks
[2] Leaderboard: https://ai4media-bench.aimultimedialab.ro/competitions/6/
[3] https://www.imageclef.org/2026/medical/mediqa-core

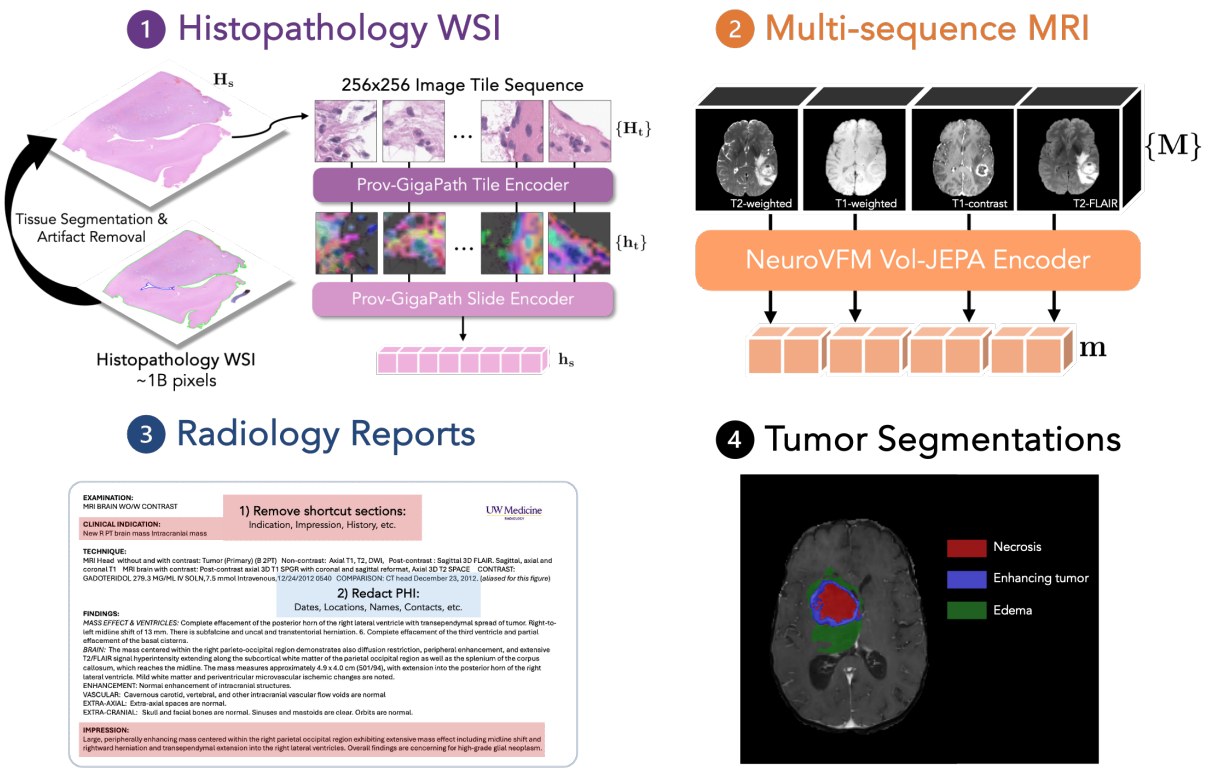


**Figure 1:** Overview of the CoRe-BT dataset.

## 3. Data Description

We used the CoRe-BT dataset [11], a new dataset that provides a multimodal platform for glioma classification by combining volumetric MRI, digital histopathology, and reports collected from a de-identified U.S. hospital cohort[4]. The dataset is designed to support multimodal tumor characterization under heterogeneous modality availability conditions. Table 1 summarizes the distribution of samples across the training, validation, and test sets under different modality availability settings.

**Table 1**
Data Distribution across Train, Validation, and Test Sets under Different Modality Settings

| Split | Both Present | Histopathology Only | MRI Only | Total |
|---|---|---|---|---|
| Train | 35 | 16 | 132 | 183 |
| Validation | 8 | 3 | 31 | 42 |
| Test #1 | 36 | 0 | 0 | 36 |
| Test #2 | 0 | 0 | 18 | 18 |
| Full Dataset | 79 | 19 | 181 | 279 |

Figure 1 presents an overview of the CoRe-BT dataset. Given the large size of the raw dataset, both raw imaging data and precomputed MRI and pathology embeddings were made available so that participants could experiment with multimodal models without needing extensive image preprocessing or high-memory GPU clusters. The following modalities and derived representations were included:

1. H&E-stained whole-slide histopathology imaging (WSI), pathology tile embeddings ($\{\mathbf{h}_t\}$) and slide-level embeddings ($\mathbf{h}_s$) extracted using GigaPath [12], and preprocessing artifacts generated with CLAM [13] (including tissue segmentation previews and patch coordinate metadata).

[4]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/KurtLabUW/CoreBT

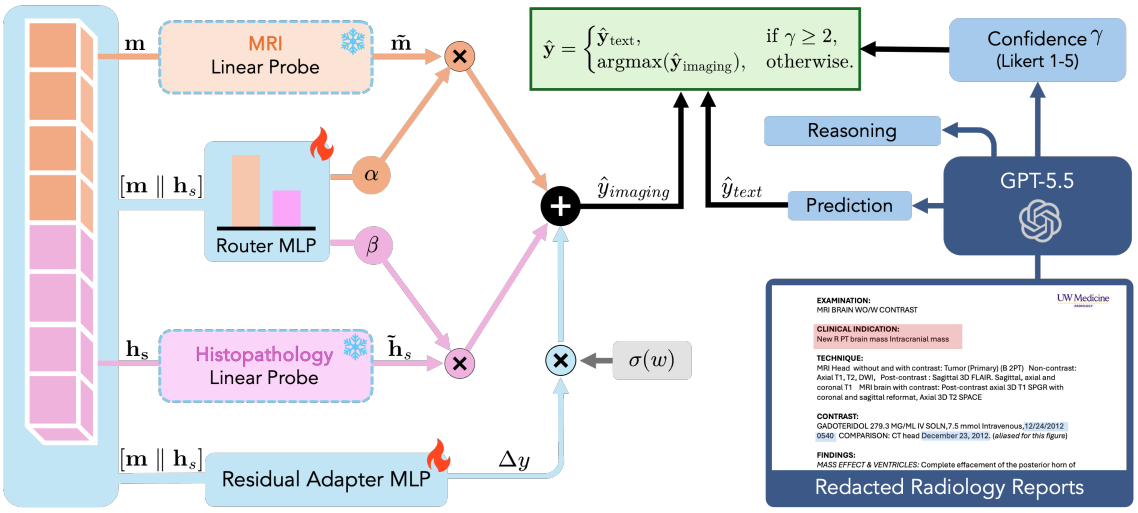


**Figure 2:** Overview of the CoRe-BT-Fusion framework.

2. Multi-sequence brain MRI volumes including T1, T2, T2-FLAIR, and T1c sequences along with global subject-level MRI embeddings ($\mathbf{m}$) extracted using NeuroVFM [14].
3. De-identified free-text radiology reports derived from the corresponding MRI studies. Radiology report sections with shortcut solutions to any classification tasks were also removed to avoid label leakage.
4. Glioma MRI segmentations annotated by two trained medical scribes following the BraTS 2023 convention [15].

MRI embeddings ($\mathbf{m}$) are generated using NeuroVFM. First, the foreground of each MRI volume $M_k$ is split into volumetric image patches and tokenized. The tokens from all available MRI series are then concatenated and encoded by the NeuroVFM student encoder, producing a variable-length sequence of embeddings with shape $N_{\text{tokens}} \times 768$, where 768 is the embedding dimension.

Histopathology WSIs are typically gigapixel in scale, making direct end-to-end processing computationally expensive. Therefore, participants are provided with embeddings extracted using Prov-GigaPath [12], a whole-slide histopathology foundation model pretrained on 1.3 billion image tiles from more than 170,000 WSIs. Slides ($H_s$) are first partitioned into $256 \times 256$ image tiles $\{H_t\}$ using the CLAM preprocessing pipeline along with a custom HSV-based artifact removal to exclude background regions and slide annotations. The tiles are then encoded using a fine-tuned DINO-v2 encoder to produce a set of tile embeddings $\{\mathbf{h}_t\}$, where each $\mathbf{h}_t = f_{\text{GigaPath\_Tile}}(H_t)$. A LongNet-adapted vision transformer aggregates information across these tile embeddings to generate the final whole-slide representation $\mathbf{h}_s = f_{\text{GigaPath\_Slide}}(\{\mathbf{h}_t\})$, capturing both local pathological structures and global signatures across the whole slide.

In addition to embeddings, the visualization artifacts include low-resolution tissue thumbnails, segmentation previews, stitched slide visualizations, and tile coordinate metadata, all used to inspect the pathology preprocessing pipeline and embedding quality.

## 4. Baseline System

Our baseline system[5], CoRe-BT-Fusion, is a multimodal fusion framework combining radiology reports, MRI, and WSI. Our approach generates independent predictions from imaging modalities and text, then combines these into a final prediction by prioritizing high-confidence text predictions. Figure 2 provides an overview of the CoRe-BT-Fusion framework [11].

[5]Code available at https://github.com/KurtLabUW/CoreBT

The imaging classification branch processes precomputed NeuroVFM MRI embeddings ($\mathbf{m}$) and GigaPath histopathology slide-level embeddings ($\mathbf{h}_s$) via two parallel pathways: a gated routing network and a residual adapter pathway. Here, the MRI study representation, $\mathbf{m}$, is defined as the mean of the sequence embeddings, $\mathbf{m} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{m}_k$, where each $\mathbf{m}_k = f_{\text{NeuroVFM}}(M_k)$ is the embedding extracted from one of the $K$ input MRI sequence volumes $M_k$. The histopathology representation $\mathbf{h}_s$ is taken from the final layer of the Prov-GigaPath slide encoder.

The routing network pathway computes a baseline prediction:

$$\mathbf{y}_{\text{base}} = \alpha\tilde{\mathbf{m}} + \beta\tilde{\mathbf{h}}_s,$$

where $\alpha$ and $\beta$ are learned routing weights. The vectors $\tilde{\mathbf{m}} = f_m(\mathbf{m})$ and $\tilde{\mathbf{h}}_s = f_h(\mathbf{h}_s)$ denote the class-logit outputs of frozen, pre-trained single-layer linear probes with LayerNorm. The routing weights are calculated using a two-layer MLP. If a modality is missing, its corresponding routing logit is masked to $-\infty$, ensuring that the following softmax operation assigns a routing weight of zero to the missing modality.

The final prediction from the imaging data $\hat{\mathbf{y}}_{\text{imaging}}$ is the sum of the gated prediction with a sigmoid-weighted residual adapter branch:

$$\hat{\mathbf{y}}_{\text{imaging}} = \mathbf{y}_{\text{base}} + \sigma(w)\Delta\mathbf{y}.$$

Here, $\Delta\mathbf{y}$ is a learned residual vector obtained by passing the concatenated features $[\mathbf{m} \parallel \mathbf{h}_s]$ through a two-layer MLP. The parameter $w$ is a learnable scalar initialized to $-4.0$, and $\sigma(\cdot)$ denotes the sigmoid function. A more detailed description of this framework is available in [11].

Text-based predictions are obtained by inputting the radiology reports into GPT-5.5 [16] prompted to produce classification outputs $\hat{\mathbf{y}}_{text}$, reasoning, and confidence scores $\gamma$ (Likert-style 1-5). The confidence threshold was selected based on validation set performance, with a threshold of $\gamma = 2$ yielding the best results among the evaluated alternatives.

For each subject, the final prediction $\hat{\mathbf{y}}$ is defined as:

$$\hat{\mathbf{y}} = \begin{cases} \hat{\mathbf{y}}_{\text{text}}, & \text{if } \gamma \geq 2, \\ \hat{\mathbf{y}}_{\text{imaging}}, & \text{otherwise.} \end{cases}$$

## 5. Evaluation Metrics

System performance is independently evaluated for the three classification tasks using the macro-averaged F1 score, precision, and recall. Macro averaging assigns equal weight to each class irrespective of class frequency, thereby emphasizing balanced performance across heterogeneous tumor categories and grading groups. For team ranking, the macro-F1 scores from the three tasks are averaged with equal weight to produce a single overall score for team comparison.

## 6. Participating Teams

MEDIQA-CORE 2026 Task 1 attracted 30 registered teams from different countries. Among them, five finalist teams submitted their runs following the challenge rules. Table 2 presents the teams that participated in the challenge. We limited the number of submitted runs to 10 runs per team.

## 7. Official Results

Table 3 presents the results on the validation set with fully multimodal and missing-modality cases. Table 4 presents the official results on test set #1 under fully multimodal and missing-modality conditions. Table 5 presents the official results on test set #2 set with MRI only cases. The results reported for the test sets were obtained after organizer-side code verification and execution of the submitted team

**Table 2**
Participating teams, papers, and code verification (availability).

| Team | Affiliation(s) | Code |
|---|---|---|
| **NeuroVietCH** [17]<br>catcd | VNU University of Engineering and Technology, Vietnam; Lausanne University Hospital, Switzerland | Yes (Public[1]) |
| **cnvntq**<br>cnvntq | NTQ, Vietnam | No |
| **UIT_Concabietbay** [18]<br>Concabietbay<br>Nguyen Minh Thoi | University of Information Technology, Vietnam National University, Vietnam | Yes (Private) |
| **DSGT** [19]<br>htruong47, cclark339 | Georgia Institute of Technology, USA | Yes (Public[4]) |
| **Brihadaranyaka** [20]<br>anubhav27 | Independent Researcher, India | Yes (Public[5]) |

[1] https://github.com/candleMind/corebt-llm-distill
[4] https://github.com/dsgt-arc/imageclef-mediqacore-2026
[5] https://github.com/AnubhavBharadwaaj/mediqatask1task2analysisnotebbok

solutions. Consequently, they may differ slightly from the publicly available leaderboard results[6], which correspond to *Test set #1* under the *fully multimodal setting* and are based on prediction files submitted by participating teams prior to code verification and organizer-side execution.

**Table 3**
Validation-set results with fully multimodal and missing-modality cases.

| Rank | Team | Run ID | Level 1 Molecular Type F1 | WHO Grade F1 | LGG vs. HGG F1 | Average Score |
|---|---|---|---|---|---|---|
| 1 | UIT_Concabietbay | 566 | **0.867** | 0.615 | **0.760** | **0.747** |
| 2 | cnvntq | 646 | 0.759 | 0.614 | 0.689 | 0.687 |
| *3* | *MEDIQA Organizers* | *552* | 0.512 | **0.685** | 0.735 | 0.644 |

On Test set #1, under the fully multimodal setting, *NeuroVietCH* achieved near-perfect performance across all tasks, obtaining an average F1 score of 0.947. Among the teams with available modality-ablation results, *DSGT* achieved the second-highest average F1 score (0.801), slightly outperforming our baseline (0.796). *UIT_Concabietbay* obtained the second-best Level 1 F1 score (0.878), whereas our baseline achieved the second-best WHO Grade F1 score (0.795) and LGG vs. HGG F1 score (0.920).

The modality-ablation results demonstrate that the relative utility of each modality varies across teams. When pathology images were excluded (MRI images + MRI reports), *Brihadaranyaka* achieved the highest average F1 score (0.746), narrowly outperforming the baseline (0.744), which achieved the best WHO Grade F1 (0.726) and LGG vs. HGG F1 (0.920). When MRI reports were excluded (MRI images + pathology images), *NeuroVietCH* maintained the highest average F1 score (0.947), matching its fully multimodal performance, whereas the remaining teams experienced only modest changes relative to the fully multimodal setting. Finally, when MRI images were excluded (pathology images + MRI reports), *NeuroVietCH* again achieved the highest average F1 score (0.916), while the baseline obtained the second-highest average F1 score (0.796) and the second-best WHO Grade F1 (0.795) and LGG vs. HGG F1 (0.920). Overall, removing MRI reports had relatively little impact on performance for several teams, whereas excluding either imaging modality generally resulted in larger performance degradation, highlighting the complementary value of MRI and pathology images.

On Test set #2, performance dropped substantially for all participating teams compared with Test set #1. Under the MRI images-only setting, participant performance remained modest, with *UIT_Concabietbay*

[6] https://ai4media-bench.aimultimedialab.ro/competitions/6/

**Table 4**
Official test-set #1 results. Teams are ranked within each condition by average macro F1 score (Avg.). The best and second-best verified results in each column are shown in bold and underlined, respectively. Scores shown in gray were not verified because the corresponding code was not submitted.

| Condition | Team | Rank | Level 1 F1 | WHO Grade F1 | LGG vs. HGG F1 | Avg. |
|---|---|---|---|---|---|---|
| Fully Multimodal: *MRI Images + MRI Reports + Pathology Images* | NeuroVietCH | 1 | **0.940** | **0.903** | **1.000** | **0.947** |
| | cnvntq | | 0.918 | 0.886 | 0.911 | 0.905 |
| | UIT_Concabietbay | 4 | 0.878 | 0.729 | 0.700 | 0.769 |
| | DSGT | 2 | 0.867 | 0.737 | 0.800 | 0.801 |
| | Brihadaranyaka | 5 | 0.837 | 0.649 | 0.800 | 0.762 |
| | *MEDIQA Organizers* | 3 | *0.672* | *0.795* | *0.920* | *0.796* |
| MRI Images + MRI Reports | NeuroVietCH | 4 | 0.600 | 0.698 | 0.705 | 0.668 |
| | UIT_Concabietbay | 3 | 0.666 | 0.611 | 0.807 | 0.695 |
| | DSGT | 5 | 0.455 | 0.610 | 0.859 | 0.642 |
| | Brihadaranyaka | 1 | **0.813** | 0.601 | 0.823 | **0.746** |
| | *MEDIQA Organizers* | 2 | *0.587* | **0.726** | **0.920** | *0.744* |
| MRI Images + Pathology Images | NeuroVietCH | 1 | **0.940** | **0.903** | **1.000** | **0.947** |
| | UIT_Concabietbay | 3 | 0.878 | 0.729 | 0.700 | 0.769 |
| | DSGT | 4 | 0.841 | 0.519 | 0.800 | 0.720 |
| | Brihadaranyaka | 2 | 0.837 | 0.653 | 0.823 | 0.771 |
| | *MEDIQA Organizers* | 5 | *0.534* | *0.564* | *0.438* | *0.512* |
| Pathology Images + MRI Reports | NeuroVietCH | 1 | **0.918** | **0.904** | **0.926** | **0.916** |
| | UIT_Concabietbay | 5 | 0.515 | 0.756 | 0.700 | 0.657 |
| | DSGT | 3 | 0.873 | 0.676 | 0.765 | 0.772 |
| | Brihadaranyaka | 4 | 0.492 | 0.682 | 0.800 | 0.658 |
| | *MEDIQA Organizers* | 2 | *0.672* | *0.795* | *0.920* | *0.796* |

**Table 5**
Official test-set #2 results. Teams are ranked within each condition by average macro F1 score (Avg.). The best and second-best verified results in each column are shown in bold and underlined, respectively.

| Condition | Team | Rank | Level 1 F1 | WHO Grade F1 | LGG vs. HGG F1 | Avg. |
|---|---|---|---|---|---|---|
| MRI Images + Reports | NeuroVietCH | 4 | 0.393 | 0.540 | 0.575 | 0.503 |
| | UIT_Concabietbay | 2 | 0.402 | 0.508 | 0.730 | 0.547 |
| | DSGT | 5 | 0.368 | 0.439 | 0.673 | 0.493 |
| | Brihadaranyaka | 3 | 0.344 | 0.517 | 0.730 | 0.530 |
| | *MEDIQA Organizers* | 1 | **0.816** | **0.716** | **0.858** | **0.797** |
| MRI Images Only | NeuroVietCH | 3 | 0.393 | **0.540** | 0.575 | 0.503 |
| | UIT_Concabietbay | 1 | **0.402** | 0.508 | **0.730** | **0.547** |
| | DSGT | 4 | 0.368 | 0.439 | 0.673 | 0.493 |
| | Brihadaranyaka | 2 | 0.344 | 0.496 | **0.730** | 0.523 |
| | *MEDIQA Organizers* | 5 | *0.373* | *0.409* | *0.433* | *0.405* |
| MRI Reports Only | NeuroVietCH | | - | - | - | - |
| | UIT_Concabietbay | 3 | 0.127 | 0.231 | 0.433 | 0.264 |
| | DSGT | 4 | 0.127 | 0.127 | 0.433 | 0.229 |
| | Brihadaranyaka | 2 | 0.597 | 0.182 | 0.510 | 0.430 |
| | *MEDIQA Organizers* | 1 | **0.808** | **0.567** | **0.858** | **0.744** |

achieving the highest average F1 score (0.547), followed by *Brihadaranyaka* (0.523). Task-specific performance was similarly limited, with *UIT_Concabietbay* obtaining the highest Level 1 F1 score (0.402), *NeuroVietCH* achieving the highest WHO Grade F1 score (0.540), and both *UIT_Concabietbay* and *Brihadaranyaka* tying for the best LGG vs. HGG F1 score (0.730). In contrast, our baseline achieved an average F1 score of 0.405 under the MRI images-only setting.

Because Test set #2 contains no histopathology images, the only difference between the MRI images + reports and MRI images-only settings is the availability of MRI reports. Participant performance changed only marginally between these settings, suggesting that most submitted systems derived

**Table 6**
Summary of the finalist team methodologies. All teams utilized the provided NeuroVFM MRI embeddings and GigaPath pathology embeddings.

| Team | Fusion Strategy | Learning Strategy | LLM Usage | Missing-Modality Handling |
|---|---|---|---|---|
| NeuroVietCH | Gated multimodal fusion with expert probes | LLM-guided teacher-student distillation | Prediction refinement and distillation | Learnable missing-pathology token |
| UIT_Concabietbay | Feature-level fusion of pooled embeddings | Classical ML ensembles, stacking, and probability blending | None | Median imputation and missingness indicators |
| DSGT | Task-specific modality weighting | Cross-validation ensemble and rule-based post-processing | Report encoding | Adaptive modality-specific gating |
| Brihadaranyaka | Ensemble of MRI, pathology, and text experts | Dirichlet-weighted multimodal ensemble | Few-shot prediction overrides | Weight renormalization after modality abstention |
| Our Baseline | Adaptive multimodal fusion | Joint training with modality-specific probes | Prediction refinement | Modality-presence masks |

little benefit from additional report information. In contrast, our baseline benefited substantially from multimodal input, achieving the highest average F1 score (0.797) with F1 scores of 0.816, 0.716, and 0.858 for the Level 1, WHO Grade, and LGG vs. HGG tasks, respectively. Overall, these results highlight the difficulty of generalizing to the second test set while demonstrating the value of effectively incorporating radiology reports when available.

## 8. System Descriptions

The finalist teams explored a diverse range of multimodal learning strategies, including task-specific fusion, classical machine-learning ensembles, multimodal expert ensembles, and LLM-assisted reasoning. Although participants were provided with both the raw MRI and pathology data as well as the generated foundation-model embeddings derived from NeuroVFM and GigaPath, all finalist teams elected to incorporate the provided embeddings into their submissions. Table 6 summarizes the key methodological characteristics of the finalist systems, highlighting their fusion strategies, learning approaches, use of LLM-based reasoning, and mechanisms for handling missing modalities. Detailed descriptions of each system are presented below:

**NeuroVietCH.** The NeuroVietCH team proposed a three-stage framework combining multimodal embedding fusion, LLM-guided report-based prediction refinement, and teacher–student distillation for brain tumor subtype classification [17]. Their first-stage model integrates NeuroVFM MRI embeddings and GigaPath histopathology embeddings through a gated fusion architecture. The framework incorporates modality-specific probes, a learnable missing-pathology token, gated logit fusion, and a residual fusion pathway. The first-stage model is designed to improve robustness when pathology is unavailable or less informative. In a second stage, baseline predictions and de-identified radiology reports are provided as structured input to a large language model for clinically grounded prediction refinement and radiological evidence extraction. Finally, the refined LLM outputs are subsequently distilled into a student model, transferring report-level reasoning into a multimodal classifier without requiring LLM inference at deployment.

**UIT_Concabietbay**. The UIT_Concabietbay team developed a multimodal tabular learning framework built on precomputed imaging embeddings rather than end-to-end training [18]. MRI embeddings are transformed into fixed-length subject representations using mean and standard-deviation pool-

ing. Pathology slide embeddings are summarized through mean pooling, standard-deviation pooling, maximum pooling, norm-based attention pooling, and robust mean pooling. The resulting modality-specific feature tables are merged using subject identifiers and enriched with modality-availability indicators. Missing numeric features are handled inside model pipelines using median imputation and missingness indicators. For each prediction target, the UIT_Concabietbay team evaluated a diverse set of machine-learning models including regularized logistic regression, PCA-logistic regression, calibrated linear models, extra trees, multilayer perceptrons, support-vector machines, random forests, and histogram gradient boosting. Model selection was based on out-of-fold macro-F1, combined with greedy probability blending, optional stacking, and conservative threshold tuning for the binary target.

**DSGT**. The DSGT team proposed a trimodal framework for brain tumor classification that integrates pre-extracted MRI embeddings from NeuroVFM, slide-level histopathology embeddings from Prov-GigaPath, and report embeddings derived using pretrained language models [19]. The team explored two trimodal fusion architectures and report encoders based on RadBERT and Llama-3.1-8B-Instruct. Instead of learning a single shared multimodal representation, the final model employs task-specific gating mechanisms that independently weight modality contributions for each classification task, including Level-1 Molecular Type, WHO Grade, and LGG versus HGG prediction. Final predictions are obtained through a five-fold cross-validation ensemble with logit averaging, followed by biologically motivated post-processing rules derived from the challenge tumor classification schema and validated on the training set.

**Brihadaranyaka.** The Brihadaranyaka team built a frozen-encoder multimodal ensemble for multi-task brain tumor classification, jointly predicting molecular subtype, WHO grade, and LGG versus HGG status [20]. Their framework combines 3D MRI features from NeuroVFM, histopathology embeddings from CONCH and GigaPath aggregated using CLAM-based clustering-constrained multiple-instance learning, and clinical text embeddings from Bio-ClinicalBERT and RadBERT. Eight task-specific prediction heads generate probability distributions, which are fused using a validation-optimized Dirichlet-weighted ensemble. The system further incorporates a few-shot reasoning module based on Claude Opus 4.7, which selectively overrides ensemble predictions via confidence-threshold gating and valid-label-tuple constraint enforcement. To address incomplete data, modality-dependent components abstain when required modalities are unavailable, with their ensemble weights renormalized accordingly.

## 9. Conclusion

This paper presented an overview of MEDIQA-CORE 2026 Task 1 on multimodal brain tumor subtype classification under incomplete-modality settings. The task introduced a clinically motivated benchmark based on the CoRe-BT dataset, combining MRI, histopathology whole-slide imaging, and radiology reports for robust brain tumor subtype classification in realistic diagnostic scenarios where modalities may be unavailable or asynchronously acquired.

Participants developed diverse multimodal approaches spanning fusion architectures, embedding-based learning, classical machine-learning ensembles, and large language model-guided reasoning and refinement. Notably, all finalist systems leveraged the provided foundation-model embeddings derived from NeuroVFM and GigaPath. Top-performing systems achieved high macro-F1 scores across molecular grouping, WHO grade prediction, and LGG versus HGG classification, demonstrating the benefits of integrating complementary radiological, pathological, and textual information. Modality-ablation analyses further highlighted the complementary contributions of MRI, pathology, and radiology reports, with multimodal fusion generally yielding the strongest overall performance. At the same time, the variability observed in WHO grade prediction suggests that fine-grained tumor grading from heterogeneous multimodal data remains challenging.

Overall, the shared task establishes a benchmark for studying multimodal medical reasoning under incomplete clinical inputs and provides a foundation for future research on robust multimodal learning, missing-modality adaptation, and clinically deployable brain tumor classification systems.

## 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. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. S. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. R. Gerstner, M. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup, S. J. Price, T. R. Raviv, S. M. S. Reza, M. T. Ryan, D. Sarikaya, L. H. Schwartz, H. Shin, J. Shotton, C. A. Silva, N. J. Sousa, N. K. Subbanna, G. Székely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. B. Ünal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, K. V. Leemput, The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Medical Imaging 34 (2015) 1993–2024. URL: https://doi.org/10.1109/TMI.2014.2377694.

[2] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, K. H. Maier-Hein, Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge, in: A. Crimi, S. Bakas, H. J. Kuijf, B. H. Menze, M. Reyes (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries - Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers, Lecture Notes in Computer Science, Springer, 2017, pp. 287–297. URL: https://doi.org/10.1007/978-3-319-75238-9_25.

[3] M. C. de Verdier, R. Saluja, L. Gagnon, D. LaBella, U. Baid, N. H. Tahon, M. Foltyn-Dumitru, J. Zhang, M. Alafif, S. Baig, K. Chang, G. D'Anna, L. Deptula, D. Gupta, M. A. Haider, A. Hussain, M. Iv, M. Kontzialis, P. Manning, F. Moodi, T. Nunes, A. Simon, N. Sollmann, D. Vu, M. Adewole, J. Albrecht, U. Anazodo, R. Chai, V. Chung, S. Faghani, K. Farahani, A. F. Kazerooni, J. E. Iglesias, F. Kofler, H. Li, M. G. Linguraru, B. H. Menze, A. W. Moawad, Y. Velichko, B. Wiestler, T. Altes, P. Basavasagar, M. Bendszus, G. Brugnara, J. Cho, Y. Dhemesh, B. K. K. Fields, F. Garrett, J. Gass, L. M. Hadjiiski, J. A. Hattangadi-Gluth, C. Hess, J. L. Houk, E. Isufi, L. J. Layfield, G. Mastorakos, J. Mongan, P. Nedelec, U. Nguyen, S. Oliva, M. W. Pease, A. Rastogi, J. Sinclair, R. X. Smith, L. P. Sugrue, J. Thacker, I. Vidic, J. E. Villanueva-Meyer, N. S. White, M. Aboian, G. M. Conte, A. M. Dale, M. R. Sabuncu, T. M. Seibert, B. Weinberg, A. H. Abayazeed, R. Y. Huang, S. Turk, A. M. Rauschecker, N. Farid, P. Vollmuth, A. Nada, S. Bakas, E. Calabrese, J. D. Rudie, The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment MRI, CoRR abs/2405.18368 (2024). URL: https://doi.org/10.48550/arXiv.2405.18368.

[4] P. Mobadersany, S. Yousefi, M. Amgad, D. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, L. A. D. Cooper, Predicting cancer outcomes from histology and genomics using convolutional networks, Proceedings of the National Academy of Sciences (PNAS) 115 (2018) E2970–E2979.

[5] R. J. Chen, M. Y. Lu, J. Wang, D. F. K. Williamson, S. J. Rodig, N. I. Lindeman, F. Mahmood, Pathomic fusion: An integrated framework for fusing histopathology and genomic features for

cancer diagnosis and prognosis, IEEE Trans. Medical Imaging 41 (2022) 757–770. URL: https://doi.org/10.1109/TMI.2020.3021387.

[6] K. Farahani, T. Kurc, S. Bakas, B. A. Bearce, J. Kalpathy-Cramer, J. Freymann, J. Saltz, E. Stahlberg, G. Zaki, M. P. Nasrallah, R. T. Shinohara, Computational precision medicine radiology-pathology challenge on brain tumor classification 2020, 2020. URL: https://doi.org/10.5281/zenodo.3718894. doi:10.5281/zenodo.3718894.

[7] Z. Wang, Z. Wu, D. Agarwal, J. Sun, Medclip: Contrastive learning from unpaired medical images and text, CoRR abs/2210.10163 (2022). URL: https://doi.org/10.48550/arXiv.2210.10163.

[8] S. Bannur, S. L. Hyland, Q. Liu, F. Pérez-García, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, A. Schwaighofer, M. Wetscherek, M. P. Lungren, A. V. Nori, J. Alvarez-Valle, O. Oktay, Learning to exploit temporal structure for biomedical vision-language processing, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, 2023, pp. 15016–15027. URL: https://doi.org/10.1109/CVPR52729.2023.01442.

[9] O. A. Adeyi, Pathology services in developing countries—the west african experience, Archives of pathology & laboratory medicine 135 (2011) 183–186.

[10] 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.

[11] J. H. Rivera, D. Low, X. Xiong, J. Ruzevick, D. D. Child, W. Yim, M. Kurt, A. Ben Abacha, Core-bt: A multimodal radiology-pathology-text benchmark for robust brain tumor typing, CoRR abs/2603.03618 (2026). URL: https://doi.org/10.48550/arXiv.2603.03618.

[12] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, H. Poon, A whole-slide foundation model for digital pathology from real-world data, Nature (2024). URL: https://www.nature.com/articles/s41586-024-07441-w.

[13] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nature Biomedical Engineering 5, 555–570 (2021). URL: https://www.nature.com/articles/s41551-020-00682-w.

[14] A. Kondepudi, A. Rao, C. Zhao, Y. Lyu, E. S. Harake, S. Banerjee, R. S. Joshi, A. Meißner, R. Hou, C. Jiang, A. Chowdury, A. Srinivasan, B. Athey, V. Gulani, A. Pandey, H. Lee, T. C. Hollon, Health system learning achieves generalist neuroimaging models, CoRR abs/2511.18640 (2025). URL: https://doi.org/10.48550/arXiv.2511.18640. doi:10.48550/ARXIV.2511.18640. arXiv:2511.18640.

[15] A. F. Kazerooni, N. Khalili, X. Liu, D. Haldar, Z. Jiang, S. M. Anwar, J. Albrecht, M. Adewole, U. Anazodo, H. Anderson, et al., The brain tumor segmentation (brats) challenge 2023: focus on pediatrics (cbtn-connect-dipgr-asnr-miccai brats-peds), ArXiv (2024) arXiv–2305.

[16] A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. Ostrow, A. Ananthram, et al., Openai gpt-5 system card, arXiv preprint arXiv:2601.03267 (2025).

[17] S.-D. Dang, D.-L. Tran, V.-H. Pham, N.-V. Chu, D. T. Le, T.-H. Do, D.-C. Can, H.-Q. Le, Neurovietch at mediqa-core-task-1 2026: Llm-guided multimodal distillation for brain tumor subtype classification, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[18] N. M. Thoi, L. N. Quang, T. B. Nguyen-Tat, Uit_concabietbay at mediqa-core-task-1 2026: Multi-

modal embedding-based ensemble learning for brain tumor subtype classification, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[19] H. T. T. Truong, C. Clark, Dsgt at mediqa-core-task-1 2026: Trimodal model fusion with task-specific gates for brain tumor subtype classification, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.
[20] A. Kumar, Brihadaranyaka at mediqa-core-task-1 2026: A frozen-encoder multimodal ensemble with llm few-shot reasoning for brain tumor subtype classification under modality sparsity, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.