

NeuroVietCH at MEDIQA-CORE-Task-1 2026: LLM-Guided Multimodal Distillation for Brain Tumor Subtype Classification

MEDIQA-CORE at CLEF 2026

Sy-Duy Dang^{1,†}, Duy-Long Tran^{1,†}, Ngoc-Vuong Chu¹, Viet-Hung Pham¹, Duyen Thi Le^{2,3},
Trung-Hieu Do^{2,4}, Duy-Cat Can^{1,5,6,*} and Hoang-Quynh Le^{1,*}

¹Faculty of Information Technology, VNU University of Engineering and Technology (VNU-UET), 144 Xuan Thuy, Hanoi, Vietnam

²Hanoi Medical University, 1 Ton That Tung, Kim Lien, Hanoi, Vietnam

³Phuc Thinh General Hospital, 19 Hoang Nghieu, Dong Tien, Thanh Hoa, Vietnam

⁴National Geriatric Hospital, 1A Phuong Mai, Phuong Mai Ward, Hanoi, Vietnam

⁵Platform of Bioinformatics, Lausanne University Hospital (CHUV), Rte de la Corniche 10, 1010 Lausanne, Switzerland

⁶Faculty of Biology and Medicine, University of Lausanne (UNIL), Unicentre, 1015 Lausanne, Switzerland

Abstract

Brain tumor subtype classification requires integrating heterogeneous clinical evidence, including MRI, histopathology, molecular interpretation, and radiology reports. MEDIQA-CORE Task 1 2026 evaluates this setting through the CoRe-BT benchmark, where systems must predict three related but non-identical targets: molecular grouping, low-grade versus high-grade glioma status, and WHO grade, under realistic missing-modality conditions. We propose NeuroVietCH, a three-stage framework combining multimodal embedding fusion, LLM-guided report-based refinement, and teacher-student distillation. First, we developed an independently implemented MRI-pathology fusion model using the provided NeuroVFM MRI embeddings and GigaPath pathology embeddings. The model uses modality-specific probes, a learnable missing-pathology token, gated logit fusion, and a residual fusion pathway to improve robustness when pathology is unavailable or less informative. Second, we used de-identified radiology reports together with first-stage model predictions as structured inputs to large language models, enabling clinically grounded prediction refinement and extraction of key radiological evidence. Third, we distilled the LLM-assisted predictions back into a student model, transferring report-level reasoning into a model-based multimodal classifier. No additional labeled patient dataset was used. Our best submitted run ranked first among 92 leaderboard submissions, achieving an average Macro-F1 of 0.985 on the official private test set, with Macro-F1 scores of 1.000 for molecular grouping, 1.000 for LGG versus HGG classification, and 0.956 for WHO grade. In additional organizer-run modality-ablation experiments, NeuroVietCH also ranked first in the fully multimodal, MRI-plus-pathology, and pathology-plus-report settings, with average Macro-F1 scores of 0.947, 0.947, and 0.916, respectively. These results suggest that LLM-guided distillation can effectively integrate clinical text reasoning with multimodal visual embeddings for robust brain tumor subtype classification. The inference code and checkpoint for the submitted system are publicly available at <https://github.com/candleMind/corebt-llm-distill>.

Keywords

Brain tumor subtype classification, multimodal learning, radiology report reasoning, teacher-student distillation, large language models, missing-modality learning, MRI-pathology fusion, MEDIQA-CORE

1. Introduction

1.1. Clinical and Task Motivation

Brain tumor typing requires integrating radiological appearance, histopathological morphology, and molecular alterations. The 2021 World Health Organization classification of central nervous system

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

[†]These authors contributed equally.

✉ 21020461@vnu.edu.vn (S. Dang); 21020547@vnu.edu.vn (D. Tran); 21020674@vnu.edu.vn (N. Chu); 23020542@vnu.edu.vn (V. Pham); 05220690@daihocyhanoi.edu.vn (D. T. Le); dotrunghieu@hmu.edu.vn (T. Do); duy-cat.can@chuv.ch (D. Can); lhquynh@vnu.edu.vn (H. Le)

ORCID 0009-0006-3466-1722 (T. Do); 0000-0002-6861-2893 (D. Can); 0000-0002-1778-0600 (H. Le)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

tumors formalized molecularly defined glioma entities, making biomarkers such as IDH mutation and 1p/19q codeletion central to diagnosis rather than supplementary descriptors [1]. In clinical practice, however, these sources of evidence are not always available at the same time. MRI is usually obtained early and provides information about tumor location, enhancement, necrosis, edema, and infiltrative behavior, while histopathological and molecular confirmation may be delayed or incomplete. Radiology reports provide an additional source of expert interpretation, often summarizing clinically meaningful imaging evidence that may not be fully captured by image embeddings alone.

MEDIQA-CORE Task 1 2026 evaluates this realistic diagnostic setting through multimodal brain tumor subtype classification under missing-modality conditions [2]. The task is based on CoRe-BT, a radiology–pathology–text benchmark containing MRI, whole-slide pathology, reports, segmentations, and pretrained MRI/pathology embeddings [3]. Participants predict three related targets: molecular grouping, low-grade versus high-grade glioma, and WHO grade. These targets are clinically connected but not interchangeable: LGG/HGG classification mainly separates WHO grade 2 from grades 3–4, whereas WHO grading requires finer discrimination, and molecular grouping reflects biological subtype. The task therefore requires models that can combine heterogeneous evidence while remaining robust when pathology is unavailable.

1.2. Technical Challenges

The task poses three main technical challenges. **First**, MRI and pathology describe tumors at different biological scales and are embedded by different foundation models, NeuroVFM for neuroimaging and GigaPath for pathology [4, 5]. Their feature spaces are therefore not naturally aligned, and simple feature concatenation may be insufficient. **Second**, histopathology may be absent at inference time, so the model must preserve a stable prediction pathway while explicitly handling missing modality information. **Third**, radiology reports contain useful clinical reasoning, but direct LLM-based prediction can be sensitive to prompt wording, model version, and input context [6]. A robust system should therefore use LLMs in a controlled way, integrating report-level reasoning without relying solely on direct LLM inference.

1.3. Approach and Contributions

We propose **NeuroVietCH**, a three-stage framework for MEDIQA-CORE Task 1. **First**, we build an independently implemented multimodal embedding model that combines MRI and pathology representations through modality-specific experts and adaptive fusion. **Second**, we use de-identified radiology reports together with first-stage model predictions as structured inputs to LLMs, producing refined labels and concise radiological evidence. **Third**, we distill the LLM-assisted predictions back into a student model, transferring report-level reasoning into a structured multimodal classifier.

The main contributions of this work are as follows:

- We develop an independent multimodal fusion system for CoRe-BT that combines MRI and pathology embeddings while explicitly handling missing pathology.
- We introduce an LLM-guided refinement stage that uses de-identified radiology reports and first-stage model predictions to generate clinically grounded teacher labels.
- We distill LLM-assisted predictions into a model-based student system, reducing reliance on direct LLM inference.
- We release the inference code and checkpoint for the submitted system at <https://github.com/candleMind/corebt-llm-distill>, enabling reproduction of the model-based inference pipeline without redistributing private CoRe-BT data.
- Our best submitted run ranked first among 92 submissions in the official MEDIQA-CORE Task 1 evaluation, achieving an average Macro-F1 score of 0.985 on the hidden private test set.

2. Related Work and Background

Multimodal Brain Tumor Classification. Brain tumor classification is inherently multimodal. MRI captures macroscopic tumor phenotype, pathology captures cellular morphology and tissue architecture, and clinical reports encode expert interpretation. MRI-based radiomics and deep learning have been widely studied for glioma grading and molecular prediction, including IDH mutation and 1p/19q codeletion [7, 8]. Computational pathology provides complementary evidence, with recent machine learning systems predicting WHO-defined glioma classes and molecularly informed subtypes from cryosection or whole-slide pathology images [9, 10]. The MEDIQA-CORE setting combines these sources of evidence in a single benchmark, requiring prediction of molecular grouping, LGG/HGG status, and WHO grade under incomplete modality availability [2, 3].

Foundation-Model Embeddings for MRI and Pathology. Foundation-model embeddings reduce the computational burden of working directly with volumetric MRI and gigapixel whole-slide images. In CoRe-BT, MRI embeddings are derived from NeuroVFM, while pathology embeddings are derived from GigaPath [3, 4, 5]. For pathology, CLAM established a practical weakly supervised framework for whole-slide learning and tissue-region selection from slide-level labels [11]. GigaPath extends this direction with a whole-slide foundation model trained on large-scale real-world pathology data [5]. Although these embeddings make challenge participation feasible on standard hardware, task-specific fusion remains necessary because MRI and pathology embeddings are produced by different encoders and represent different biological scales.

Missing-Modality Multimodal Learning. Missing modalities are common in clinical machine learning because not every patient has the same tests, imaging sequences, or tissue data available. In medical imaging, HeMIS addressed hetero-modal segmentation by learning modality-specific representations that can be aggregated when only a subset of modalities is available [12]. EmbraceNet similarly proposed robust multimodal fusion under partial modality absence [13]. More recent work emphasizes the importance of separating shared and modality-specific information for missing-modality learning [14, 15]. These ideas motivate our use of modality-specific experts, explicit missing-pathology handling, and adaptive fusion rather than naive concatenation.

LLMs for Clinical Report Reasoning and Distillation. Large language models can process clinical text and support medical reasoning, but their outputs may vary with prompts, model versions, and context [6, 16]. In radiology, LLMs have been studied for information extraction from reports, where they can identify clinically meaningful findings from narrative text [17]. In this work, we do not treat LLMs as diagnostic replacements. Instead, we use them as report-reasoning teachers that combine de-identified radiology reports with structured model predictions.

Teacher-student distillation provides a mechanism for transferring knowledge from a stronger or more complex teacher into a deployable student model [18]. Recent work further shows that LLM-generated rationales can serve as intermediate supervision for smaller task-specific models [19]. Our framework follows this principle by using LLM-assisted predictions as teacher signals and distilling them back into a structured multimodal classifier.

3. Dataset, Tasks, and Evaluation Setting

CoRe-BT Dataset. We used the CoRe-BT dataset released for MEDIQA-CORE Task 1 2026 [3]. CoRe-BT is a subject-level multimodal benchmark for brain tumor typing, combining MRI, whole-slide histopathology, radiology reports, tumor segmentations, and pretrained modality embeddings. The MRI data include T1, contrast-enhanced T1, T2, and FLAIR sequences, with tumor segmentations following the BraTS-style convention. The pathology data include whole-slide images and derived GigaPath embeddings, while MRI embeddings are provided from NeuroVFM. These pretrained embeddings allow

Table 1

Prediction targets and label definitions in MEDIQA-CORE Task 1.

Target	Label	Class definition	Clinical meaning
Molecular grouping	0	GBM, IDH-wildtype, CNS WHO grade 4	Aggressive adult-type diffuse glioma, defined by IDH-wildtype status with histologic or molecular features of glioblastoma.
	1	IDH-mutant adult-type diffuse glioma	Includes astrocytoma, IDH-mutant, and oligodendroglioma, IDH-mutant and 1p/19q-codeleted.
	2	Non-diffuse / non-adult type diffuse gliomas	Includes pediatric-type diffuse gliomas, circumscribed astrocytic gliomas, glioneuronal tumors, and other CNS tumor types.
LGG vs HGG	0	Low-grade glioma / lower-grade tumor	Less aggressive tumors, generally corresponding to CNS WHO grade 1–2.
	1	High-grade glioma	More aggressive tumors, generally corresponding to CNS WHO grade 3–4.
WHO grade	0	CNS WHO grade 1–2	Low-grade or less aggressive tumor group.
	1	CNS WHO grade 3	Intermediate high-grade tumor group, more aggressive than grade 2 but not meeting criteria for grade 4.
	2	CNS WHO grade 4	Highest-grade tumor group, typically associated with aggressive histology and/or defining molecular alterations.

participants to develop multimodal classifiers without processing all raw volumetric MRI or gigapixel whole-slide images directly.

The dataset reflects an incomplete-modality clinical setting. MRI information is available for all subjects, whereas pathology may be unavailable for some cases at inference time. No case can be excluded because of missing pathology, requiring systems to remain functional under incomplete multimodal input.

Prediction Targets. MEDIQA-CORE Task 1 evaluates three related prediction targets: molecular grouping, LGG/HGG status, and WHO grade. The corresponding submission columns are `level1_pred`, `lgghgg_pred`, and `who_grade_pred`, respectively. Table 1 summarizes the label space and clinical interpretation of each class.

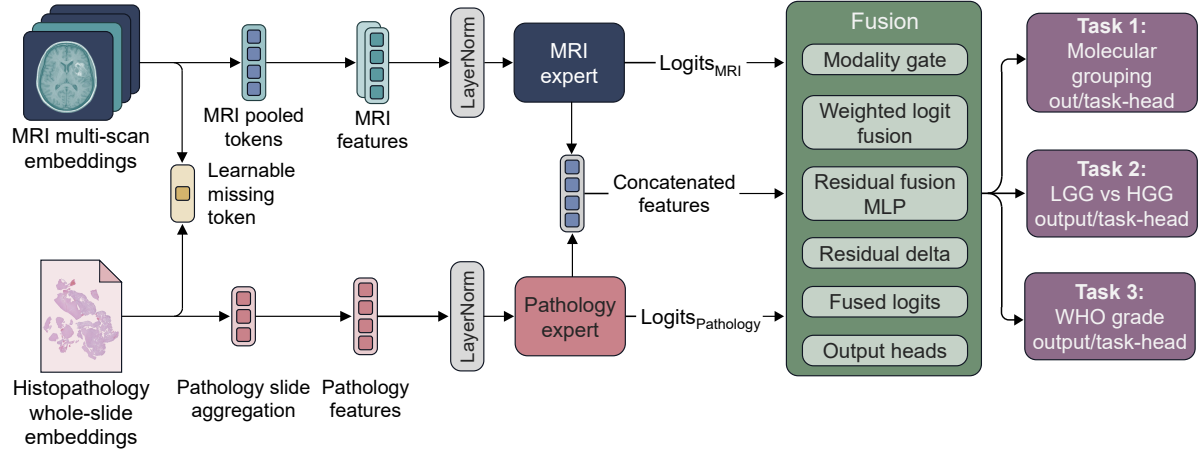
Although the three targets are related, they are not interchangeable. LGG/HGG classification largely separates WHO grade 2 from grades 3–4, whereas WHO grading requires finer discrimination, particularly between grades 3 and 4. Molecular grouping reflects biological subtype and may require evidence beyond histological aggressiveness alone. This label structure motivated our use of both multimodal embedding-based prediction and report-guided reasoning.

Evaluation Setting. Submissions were evaluated on the official hidden private test set through the MEDIQA-CORE Task 1 evaluation server.¹ Private test labels were not available to participants. The organizers reported Macro-F1, precision, and recall for each task, with the main leaderboard score computed as the average of the three task-level Macro-F1 scores. The leaderboard contained 92 total submissions. Our best NeuroVietCH submission, Run ID 1066, ranked first in the official main results.

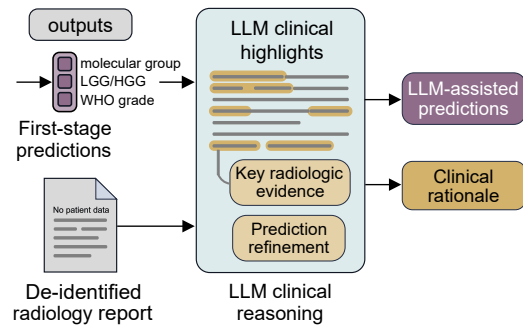
After the main evaluation, the organizers performed additional experiments on a separate test setting to assess system robustness under different modality-availability scenarios. The organizer-run analysis therefore evaluated finalist systems under four settings: fully multimodal input, MRI images plus MRI reports, MRI images plus pathology images, and pathology images plus MRI reports. These experiments provide complementary evidence about missing-modality behavior and the relative contribution of each modality, but they are separate from the official main leaderboard.

¹<https://ai4media-bench.aimultimedialab.ro/competitions/6/>

A First-stage multimodal model



B LLM-guided prediction refinement



C Teacher-student distillation

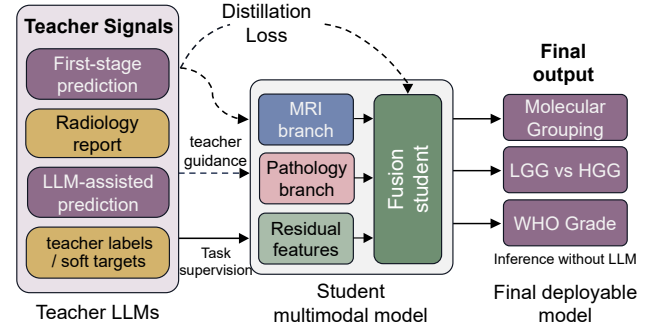


Figure 1: Overall architecture of the NeuroVietCH system. (A) The first-stage multimodal model integrates MRI and pathology embeddings using modality-specific experts, missing-pathology handling, gated logit fusion, and residual fusion. (B) De-identified radiology reports and first-stage predictions are used for LLM-guided prediction refinement and evidence extraction. (C) LLM-assisted predictions are used as teacher labels or soft targets for student-model distillation, producing final predictions without requiring LLM inference.

4. Method

Our system consists of three stages: an embedding-based multimodal classifier, an LLM-guided refinement module using de-identified radiology reports, and a teacher–student distillation stage that transfers LLM-assisted predictions back into a deployable student model. Figure 1 summarizes the overall NeuroVietCH framework.

4.1. Feature Preparation

MRI Features. We used the provided padded T1CE MRI embeddings as the primary MRI representation. Preliminary experiments showed that T1CE and T2-FLAIR embeddings were more stable than T2W embeddings. We selected T1CE to reduce inconsistencies introduced by variable availability across MRI sequences. Given a subject-level sequence of MRI tokens, we applied mean pooling to obtain a 768-dimensional MRI feature vector $\mathbf{x}_m$.

Pathology Features. For pathology, we used the provided GigaPath-based slide embeddings. When multiple slides were available for a subject, we mean-aggregated slide embeddings to obtain a single 768-dimensional pathology feature vector $\mathbf{x}_p$. No raw WSI preprocessing was performed for the submitted system.

Missing Pathology. Because histopathology was unavailable for some cases, we introduced a learnable missing-pathology token $\mathbf{x}_{\text{miss}}$. When pathology embeddings were absent, $\mathbf{x}_{\text{miss}}$ replaced $\mathbf{x}_p$. This preserved a fixed fusion architecture while allowing the model to learn a task-specific representation of missing pathology.

4.2. Multimodal Fusion Model

Unimodal Expert Probes. We first trained MRI and pathology probes independently. The MRI probe f_m maps the MRI feature vector to modality-specific logits,

$$\mathbf{z}_m = f_m(\mathbf{x}_m), \quad (1)$$

and the pathology probe f_p maps the pathology feature vector, or the missing-pathology token, to pathology-specific logits,

$$\mathbf{z}_p = f_p(\mathbf{x}_p). \quad (2)$$

After unimodal training, both probes were frozen and reused as modality-specific experts in the fusion model.

Gated Logit Fusion. The fusion model learns modality weights from the concatenated feature vector $[\mathbf{x}_m; \mathbf{x}_p]$. Let $\mathbf{g} = (g_m, g_p)$ denote the two-dimensional modality gate:

$$\mathbf{g} = \text{softmax}(h([\mathbf{x}_m; \mathbf{x}_p])), \quad (3)$$

where h is a small gating network. The gated logit prediction is computed as

$$\mathbf{z}_{\text{gate}} = g_m \mathbf{z}_m + g_p \mathbf{z}_p. \quad (4)$$

This mechanism allows the model to adaptively weight MRI and pathology predictions, especially when pathology is absent or less informative.

Residual Fusion Pathway. In parallel with gated logit fusion, we used a residual fusion MLP r to model feature-level interactions between modalities:

$$\Delta \mathbf{z} = r([\mathbf{x}_m; \mathbf{x}_p]). \quad (5)$$

The final fused logits are

$$\mathbf{z}_{\text{fused}} = \mathbf{z}_{\text{gate}} + \Delta \mathbf{z}. \quad (6)$$

The residual pathway allows the model to correct the late-fused unimodal prediction using joint MRI–pathology feature information.

Task-Specific Outputs. The fused logits were used to generate predictions for the three task outputs: molecular grouping, LGG/HGG status, and WHO grade. In the submitted systems, task-specific classifiers were trained according to the corresponding label space.

4.3. Training Strategy

The first-stage multimodal model was trained using a staged expert-fusion strategy, consistent with Panel A of Figure 1. We first trained the MRI branch and pathology branch as independent unimodal experts. The MRI expert was trained from the mean-pooled MRI feature vector $\mathbf{x}_m$, while the pathology expert was trained from the subject-level pathology feature vector $\mathbf{x}_p$. After unimodal training, the selected MRI and pathology experts were frozen and reused to provide modality-specific logits $\mathbf{z}_m$ and $\mathbf{z}_p$ during fusion training.

The fusion stage was then trained on top of the frozen experts. It received both the expert logits and the concatenated feature representation $[\mathbf{x}_m; \mathbf{x}_p]$. The modality gate learned how much to rely on each expert, while the residual fusion MLP learned an additional correction from the joint MRI–pathology feature representation. The resulting fused logits were passed to the task-specific output heads for molecular grouping, LGG/HGG status, and WHO grade.

Because the official score is based on Macro-F1 and the label distribution is imbalanced, we used class-balanced sampling during training. Model selection was guided by validation Macro-F1 and cross-validation behavior rather than by a single checkpoint. For the MRI expert, we selected a checkpoint with consistently strong train/validation Macro-F1. For the pathology expert, we selected a later high-performing checkpoint that produced better downstream fusion behavior than the earliest standalone validation peak.

To reduce sensitivity to short-term validation fluctuations, we used EMA-based checkpoint selection for the fusion model. After selecting the training configuration, the final Task 1 fusion model was trained using the selected frozen experts, class-balanced sampling, gated logit fusion, residual feature fusion, and task-specific output heads.

4.4. LLM-Guided Prediction Refinement

The embedding-based fusion model provides a structured multimodal prior, while the radiology report contains expert narrative evidence. We therefore used the first-stage model prediction and the de-identified radiology report jointly as input to LLMs for clinically grounded prediction refinement.

For each subject, the LLM received the task definitions, label mappings, first-stage predictions for the three tasks, and the corresponding de-identified radiology report. The prompt instructed the LLM to reason from medical and radiological evidence, treat the model prediction as a prior rather than ground truth, and check consistency between LGG/HGG status and WHO grade. The complete prompt template is provided in Appendix A.

We evaluated two commercial LLMs, referred to as ChatGPT and Gemini in this report. They were used only for prediction refinement and teacher generation. Only de-identified radiology reports and model predictions were provided to the LLMs. No patient identifiers, raw MRI files, whole-slide pathology images, or private metadata were uploaded.

The LLMs returned structured predictions for molecular grouping, LGG/HGG status, and WHO grade, together with concise evidence summaries. The structured predictions were used for submitted runs and teacher–student distillation. The evidence summaries were used to audit whether label changes were supported by plausible radiological findings.

4.5. Teacher–Student Distillation

The final stage transfers the LLM-refined predictions into a student multimodal model, as shown in Panel C of Figure 1. The teacher signal was constructed from the first-stage prediction, the de-identified radiology report, and the LLM-assisted prediction. We considered two teacher variants, one generated with ChatGPT and one generated with Gemini. The best official submission used the ChatGPT-guided teacher.

The student followed the same multimodal structure as the first-stage fusion model. It received the MRI branch, the pathology branch, and the residual feature pathway, and produced predictions through the fusion student module. Teacher labels and soft targets were used only as training supervision and were not provided as input to the student during inference. When the LLM returned a confidence score, we converted it into a task-level soft target distribution by assigning the reported confidence to the selected class and distributing the remaining probability mass uniformly across the other classes.

Let $t \in \{1, 2, 3\}$ index the three prediction tasks. For task t , the teacher provides a hard label $y_T^{(t)}$ and a soft target distribution $\mathbf{q}_T^{(t)}$, while the student outputs a probability distribution $\mathbf{p}_S^{(t)}$. The distillation

objective combines hard-label and soft-target supervision:

$$\mathcal{L}_{\text{distill}} = \sum_{t=1}^3 \left[\alpha \text{CE} \left(y_{\text{T}}^{(t)}, \mathbf{p}_{\text{S}}^{(t)} \right) + (1 - \alpha) \text{KL} \left(\mathbf{q}_{\text{T}}^{(t)} \parallel \mathbf{p}_{\text{S}}^{(t)} \right) \right], \quad (7)$$

where α is a scalar weight controlling the balance between hard teacher labels and soft teacher targets.

The final training objective combines supervised learning from the available task labels with LLM-guided distillation:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{distill}}, \quad (8)$$

where λ is a scalar weight controlling the contribution of teacher supervision. Distillation was performed using the available training labels together with LLM-generated teacher targets for the corresponding training instances.

After distillation, the final deployable model produces molecular grouping, LGG/HGG status, and WHO grade predictions directly from the multimodal embeddings. It does not require LLM calls at inference time. This design uses the LLM as a training-time report-reasoning teacher, while preserving a conventional model-based prediction pipeline for the final submission.

Importantly, the final distilled student does not consume radiology reports as an input feature during inference. Radiology reports are used during LLM-guided teacher generation, but the deployed student predicts from the available embedding-based visual modalities. Consequently, in organizer-run modality-ablation settings where MRI reports are listed as available, the report modality should be interpreted as available to the benchmark setting, not as an inference-time input used by the NeuroVietCH student model.

5. Experiments and Results

5.1. Run Configurations and External Resources

We evaluated a series of submitted variants to separate the contribution of the first-stage multimodal model, report-based LLM refinement, and teacher-student distillation. Table 2 summarizes the main system configurations.

Table 2

Main NeuroVietCH run configurations evaluated on the official private test set.

Run configuration	Emb.	Report	LLM	Student	Purpose
First-stage model only	Yes	No	No	No	Embedding-only reference.
Gemini only	No	Yes	Gemini	No	Report-only LLM reference.
ChatGPT only	No	Yes	ChatGPT	No	Stronger report-only LLM reference.
Gemini-ChatGPT voting	No	Yes	Both	No	Naive LLM ensemble reference.
First-stage model + Gemini	Yes	Yes	Gemini	No	Model-prior LLM refinement.
First-stage model + ChatGPT	Yes	Yes	ChatGPT	No	Model-prior LLM refinement.
Gemini-guided distilled student	Yes	Yes	Gemini	Yes	Distillation reference.
ChatGPT-guided distilled student	Yes	Yes	ChatGPT	Yes	Final submitted system.

We did not use any additional labeled patient datasets beyond the data released by the organizers. The provided NeuroVFM MRI embeddings and GigaPath pathology embeddings were used as input representations. The only external models used beyond the provided embeddings were commercial LLMs used for report-guided prediction refinement and teacher generation. Only de-identified radiology reports and model predictions were provided to the LLMs; no raw MRI files, whole-slide pathology images, identifiable patient information, or private metadata were uploaded.

The inference code and checkpoint for the submitted system are publicly available at <https://github.com/candleMind/corebt-llm-distill>. The repository supports reproduction of the model-based inference

Table 3

Official MEDIQA-CORE Task 1 main results. Scores are Macro-F1 for each task and their average.

Rank	Team	Run	Molecular	LGG/HGG	WHO grade	Average
1	NeuroVietCH (our model)	1066	1.000	1.000	0.956	0.985
2	cnvntq	1049	0.918	0.911	0.886	0.905
3	UIT_Concabietybay	728	0.900	0.839	0.858	0.866
4	DSGT	949	0.915	0.859	0.697	0.824
5	Brihadaranyaka	1039	0.540	0.958	0.772	0.757
6	cclark339	986	0.792	0.765	0.679	0.745
–	MEDIQA Organizers	1185	0.539	0.920	0.726	0.728

pipeline but does not include the private CoRe-BT dataset, raw MRI volumes, whole-slide pathology images, radiology reports, or other restricted challenge materials. Additional training code and extended experimental analyses may be released in a follow-up detailed study.

5.2. Official Leaderboard Performance

Table 3 reports the official MEDIQA-CORE Task 1 main results. The evaluation was performed on the hidden private test set, and private labels were not available to participants. The leaderboard score was computed as the average of the three task-level Macro-F1 scores.

NeuroVietCH ranked first among the finalist teams and achieved the highest official average score. Our best submission, Run ID 1066, obtained Macro-F1 scores of 1.000 for molecular grouping, 1.000 for LGG/HGG classification, and 0.956 for WHO grade, yielding an average Macro-F1 of 0.985. The leaderboard contained 92 total submissions. Compared with the second-ranked system, our best run improved the average Macro-F1 by 0.080, with the largest visible gains in molecular grouping and LGG/HGG classification.

5.3. Organizer-Run Modality-Ablation Analysis

Table 4 reports the organizer-run modality-ablation results. For NeuroVietCH, “MRI reports” in the organizer-defined settings should be interpreted only as an available modality in the evaluation scenario, since the final distilled student does not use reports as direct inference-time input. Reports were only used during LLM-guided teacher generation and distillation.

NeuroVietCH ranked first in three of the four organizer-run modality settings. In the fully multimodal setting, NeuroVietCH achieved an average Macro-F1 of 0.947, with Macro-F1 scores of 0.940 for molecular grouping, 1.000 for LGG/HGG classification, and 0.903 for WHO grade. The same score was obtained in the MRI-plus-pathology setting, which is consistent with the fact that the final distilled student does not directly consume radiology reports at inference time. This result suggests that the MRI and pathology embedding pathways were the primary inference-time drivers of the final student.

When pathology was unavailable, performance decreased substantially. In the MRI-images-plus-reports setting, NeuroVietCH achieved an average Macro-F1 of 0.668. Because the final student did not use reports as direct inference input, this setting effectively tested the behavior of the visual model when pathology information was removed. The decrease indicates that pathology embeddings were an important contributor to the final predictions, especially for molecular grouping.

In the pathology-images-plus-reports setting, NeuroVietCH achieved an average Macro-F1 of 0.916 and ranked first. Again, because reports were not used directly by the final student, this result primarily reflects the strength of the pathology pathway under reduced imaging availability. The high performance in this setting suggests that pathology embeddings carried strong diagnostic signal for molecular grouping and WHO grade, while the lower performance in the MRI-only visual setting highlights the complementary role of pathology in the submitted system.

Table 4

Organizer-run modality-ablation results on an additional test setting. Scores are Macro-F1 for each task and their average.

Available modalities [‡]	Team	Molecular	LGG/HGG	WHO grade	Average
MRI images + MRI reports [†] + pathology images	NeuroVietCH	0.940	1.000	0.903	0.947
	DSGT	0.867	0.800	0.737	0.801
	MEDIQA Organizers	0.672	0.920	0.795	0.796
	UIT_Concabetbay	0.878	0.700	0.729	0.769
	Brihadaranyaka	0.837	0.800	0.649	0.762
MRI images + MRI reports [†]	Brihadaranyaka	0.813	0.823	0.601	0.746
	MEDIQA Organizers	0.587	0.920	0.726	0.744
	UIT_Concabetbay	0.666	0.807	0.611	0.695
	NeuroVietCH	0.600	0.705	0.698	0.668
	DSGT	0.455	0.859	0.610	0.642
MRI images + pathology images	NeuroVietCH	0.940	1.000	0.903	0.947
	Brihadaranyaka	0.837	0.823	0.653	0.771
	UIT_Concabetbay	0.878	0.700	0.729	0.769
	DSGT	0.841	0.800	0.519	0.720
	MEDIQA Organizers	0.534	0.438	0.564	0.512
Pathology images + MRI reports [†]	NeuroVietCH	0.918	0.926	0.904	0.916
	MEDIQA Organizers	0.672	0.920	0.795	0.796
	DSGT	0.873	0.765	0.676	0.772
	Brihadaranyaka	0.492	0.800	0.682	0.658
	UIT_Concabetbay	0.515	0.700	0.756	0.657

[‡]For NeuroVietCH, “MRI images” and “pathology images” refer to the provided NeuroVFM and GigaPath embeddings used by the submitted model. [†]Radiology reports were not direct inference-time inputs to the final distilled student; they were used only during LLM-guided teacher generation.

Overall, the additional organizer-run experiments refine the interpretation of our system. The official main leaderboard demonstrates strong performance when all modalities are available, while the modality-ablation analysis shows that NeuroVietCH is strongest when pathology embeddings are available. The results also clarify that the LLM-guided report reasoning primarily improved the student through training-time teacher signals, rather than through direct report-based inference.

5.4. Comparison of NeuroVietCH Variants

In contrast to the organizer-run modality-ablation analysis above, this subsection compares our own submitted variants on the original official private test set. Table 5 compares our submitted variants on the official private test set. These results provide insight into how each stage of the pipeline contributed to the final system.

The first-stage embedding model alone achieved an average Macro-F1 of 0.626. Its strongest component was LGG/HGG classification, with Macro-F1 of 0.807, suggesting that the MRI–pathology embedding pathway captured useful low-grade versus high-grade signal. However, performance was lower for molecular grouping and WHO grade, indicating that visual embeddings alone were insufficient for the full diagnostic label space in this setting.

Text-only LLM prediction improved over the first-stage embedding model, especially for ChatGPT. ChatGPT-only achieved an average Macro-F1 of 0.814, compared with 0.644 for Gemini-only. The strongest ChatGPT-only component was LGG/HGG classification, with Macro-F1 of 0.911. This is clinically plausible because radiology reports often contain evidence of aggressiveness, including enhancement, necrosis, edema, and mass effect, which can support low-grade versus high-grade separation.

Table 5

Official private-test performance of NeuroVietCH submitted variants. Scores are Macro-F1 for each task and their average.

Run	Configuration	Molecular	LGG/HGG	WHO grade	Average
1066	ChatGPT-guided distilled student	1.000	1.000	0.956	0.985
1065	Gemini-guided distilled student	1.000	0.958	0.904	0.954
1062	First-stage model + ChatGPT refinement, run 1	1.000	0.911	0.886	0.932
1063	First-stage model + ChatGPT refinement, run 2	1.000	0.920	0.877	0.932
1064	First-stage model + Gemini refinement	1.000	0.874	0.860	0.911
733	ChatGPT only	0.741	0.911	0.791	0.814
740	Gemini-ChatGPT voting	0.729	0.807	0.717	0.751
734	Gemini only	0.567	0.723	0.643	0.644
891	First-stage model only	0.520	0.807	0.552	0.626

Providing first-stage model predictions to the LLMs substantially improved performance. The first-stage model plus ChatGPT refinement reached an average Macro-F1 of 0.932 across two runs, while the corresponding Gemini refinement reached 0.911. This indicates that the first-stage prediction acted as a useful structured prior: the LLM was not required to infer all labels from the report alone, but could reconcile the model prediction with radiological evidence. The improvement was particularly clear for molecular grouping, where all model-assisted LLM runs achieved 1.000 Macro-F1.

Naive voting between Gemini and ChatGPT did not improve performance. The voting run achieved an average Macro-F1 of 0.751, below ChatGPT-only and below model-assisted LLM refinement. This suggests that disagreements between LLMs were not random errors that could be eliminated by simple ensembling. Instead, LLM outputs required a structured prior and a stabilizing learning step to become useful for the final system.

The strongest performance was obtained after distillation. ChatGPT-guided distillation improved the average Macro-F1 from 0.932 for direct ChatGPT refinement to 0.985 for the distilled student. Relative to the first ChatGPT-refinement run, LGG/HGG Macro-F1 increased from 0.911 to 1.000, and WHO grade Macro-F1 increased from 0.886 to 0.956. Gemini-guided distillation also improved over direct Gemini refinement, increasing the average score from 0.911 to 0.954. These results support the central design of our system: LLM-assisted report reasoning is most effective when transferred into a student model rather than used only as a direct prediction mechanism.

5.5. Task-Level Interpretation

Across submitted variants, molecular grouping showed a sharp transition from weak first-stage performance to perfect performance after model-assisted LLM refinement. This suggests that radiology-report evidence and LLM reasoning were particularly helpful for distinguishing broad molecular groups when combined with the first-stage model prior. However, molecular grouping should not be interpreted as purely radiological; the result reflects the combined effect of structured model prediction, report reasoning, and distillation.

LGG/HGG classification was relatively strong even for the first-stage model and ChatGPT-only variants. This is clinically plausible because the low-grade versus high-grade distinction is often reflected in imaging descriptors such as enhancement, necrosis, edema, and aggressive mass effect. The best distilled system achieved perfect Macro-F1 for this task, indicating that the final student preserved both the visual prior and the report-derived aggressiveness signal.

WHO grade remained the most challenging target. Even in the best run, WHO grade Macro-F1 was 0.956, lower than molecular grouping and LGG/HGG classification. This is consistent with the clinical difficulty of separating grade 3 from grade 4 disease using incomplete multimodal evidence. Unlike LGG/HGG classification, WHO grading requires finer discrimination, and grade 4 assignment may depend on histological or molecular features that are not always explicit in imaging reports.

Overall, the submitted-run comparison shows that the final system benefited from the interaction of three components rather than from any single component alone. The first-stage model provided a multimodal embedding prior, the LLM used the radiology report to refine this prior, and distillation stabilized the refined predictions in a deployable student model. This explains why the best result was achieved by ChatGPT-guided distillation, rather than by the embedding model alone, LLM-only prediction, or naive voting.

6. Discussion and Conclusion

Why the Framework Worked. The official results suggest that the strongest performance came from combining complementary sources of evidence rather than relying on a single modality or model. The first-stage MRI–pathology model provided a structured multimodal prior from visual embeddings. The radiology report provided expert narrative evidence about tumor appearance, aggressiveness, location, and diagnostic impression. The LLM-guided refinement stage connected these two sources by using report-level reasoning to reassess the first-stage predictions. Finally, teacher–student distillation stabilized the LLM-refined outputs in a model-based pipeline, reducing dependence on direct LLM inference at submission time.

This interaction explains why the best result was not achieved by the embedding model alone, by LLM-only prediction, or by naive voting. The first-stage model alone captured useful LGG/HGG signal but was insufficient for molecular grouping and WHO grade. LLM-only prediction improved performance, especially for report-driven LGG/HGG reasoning, but remained sensitive to the specific model used. Model-assisted LLM refinement performed better because the first-stage prediction acted as a structured prior. Distillation then transferred these refined decisions into the student model, producing the strongest and most stable submitted system.

Clinical Interpretation. Radiology reports often encode findings such as contrast enhancement, necrosis, edema, infiltrative growth, multifocality, tumor location, and mass effect. These features can support prediction of tumor aggressiveness and may provide indirect evidence for subtype or grade. However, they are not definitive substitutes for histopathological and molecular diagnosis. The proposed system should therefore be interpreted as a competition system and a potential decision-support approach, not as a replacement for standard diagnostic workflows.

The task-level results are clinically plausible. LGG/HGG classification reached perfect Macro-F1 in the best submitted run, consistent with the fact that low-grade versus high-grade separation is often reflected in imaging and report-level descriptors of aggressiveness. WHO grade remained more difficult, with a best Macro-F1 of 0.956, because distinguishing grade 3 from grade 4 may require finer histological or molecular evidence. This supports the need for multimodal systems that can incorporate imaging, pathology, reports, and molecular context when available.

Missing-Modality Robustness. A key objective of MEDIQA-CORE Task 1 was to evaluate systems under realistic incomplete-modality conditions. However, the original official private test set contained only fully multimodal cases. The additional organizer-run modality-ablation analysis therefore provides important complementary evidence about robustness when different modalities are removed.

The results show that NeuroVietCH was strongest when pathology embeddings were available. The system achieved an average Macro-F1 of 0.947 in both the fully multimodal and MRI-plus-pathology settings, and 0.916 in the pathology-plus-report setting. Because the final distilled student does not use radiology reports as inference-time input, the identical fully multimodal and MRI-plus-pathology results indicate that direct report input did not contribute to the final prediction pathway. Instead, report information was incorporated indirectly during training through LLM-guided teacher generation and distillation.

Performance decreased to 0.668 in the MRI-images-plus-reports setting, where pathology was removed. For NeuroVietCH, this setting effectively evaluates the MRI pathway without direct use of the

report modality. The drop suggests that pathology embeddings were a major contributor to molecular grouping and WHO grade prediction in the final system. This is clinically plausible, since histopathology provides cellular and architectural evidence more directly related to tumor type and grade than MRI appearance alone.

The staged expert-fusion design nevertheless supports incomplete-input inference. The learnable missing-pathology token preserves a fixed architecture when histopathology is unavailable, while the modality gate allows the model to adapt the relative contribution of MRI and pathology experts. The additional organizer-run results therefore refine the interpretation of our approach: NeuroVietCH is robust and highly competitive under pathology-available settings, but MRI-only visual inference remains substantially harder and should be a focus of future work.

LLMs as Teachers Rather Than Final Diagnostic Systems. A central design choice was to use LLMs as teacher and refinement modules rather than as standalone diagnostic systems. This reduced dependence on direct LLM inference and made the final prediction pipeline more compatible with standard machine-learning deployment. The LLMs were used to read de-identified radiology reports, compare report evidence with first-stage model predictions, and generate structured teacher signals. The final distilled student then produced predictions from multimodal embeddings without requiring LLM calls at inference time.

The submitted variants show why this design was necessary. ChatGPT outperformed Gemini in this setting, indicating that LLM-guided refinement depended on the specific model. Naive voting between Gemini and ChatGPT was not reliable, suggesting that LLM disagreements should not be treated as independent random errors. Prompting choices and model versions may also affect reproducibility. Distillation provided a more controlled way to use LLM reasoning by transferring refined predictions into a fixed student model.

Limitations and Future Work. Several limitations should be considered. First, private-test labels were hidden, so detailed case-level error analysis was not possible. Second, the official results are leaderboard-based and should be interpreted within the shared-task setting. Third, the use of commercial LLMs affects exact reproducibility because model behavior may vary across versions, prompts, and service updates. Fourth, distillation quality depends on teacher quality, as shown by the performance difference between ChatGPT-guided and Gemini-guided distillation. Fifth, the system has not yet been externally validated beyond the challenge test set. Finally, report-based reasoning reflects radiologist interpretation and cannot fully replace direct histopathological or molecular evidence.

Future work will focus on expanded ablation studies, external validation, and more systematic analysis of modality contribution. We also plan to study LLM teacher quality more carefully, including prompt stability, disagreement patterns, and the relationship between evidence summaries and label changes. In addition, we will further investigate multimodal alignment between MRI and pathology embeddings. The additional modality-ablation results show that MRI-only visual inference remains the weakest setting for NeuroVietCH, motivating future work on stronger MRI encoders, report-aware student models, and improved cross-modal distillation when pathology is unavailable.

Conclusion. NeuroVietCH achieved first place in MEDIQA-CORE Task 1 2026 using an LLM-guided multimodal distillation framework. The method combines MRI/pathology embedding fusion, radiology-report reasoning for teacher generation, and teacher–student distillation. Our best submitted run achieved an average Macro-F1 of 0.985 on the official private test set, ranking first among 92 leaderboard submissions. In additional organizer-run modality-ablation experiments, NeuroVietCH ranked first in the fully multimodal, MRI-plus-pathology, and pathology-plus-report settings, with average Macro-F1 scores of 0.947, 0.947, and 0.916, respectively. These results suggest that LLM-guided distillation is a promising strategy for transferring clinical report reasoning into robust multimodal models, while also showing that MRI-only inference without pathology remains an important challenge for future work.

Acknowledgments

The authors thank the MEDIQA-CORE Task 1 organizers for designing and coordinating the shared task, preparing the CoRe-BT benchmark, maintaining the evaluation platform, and providing the official results. The authors also thank the reviewers for their time, constructive feedback, and suggestions, which helped improve the clarity and presentation of this work.

Declaration on Generative AI

During the preparation of this manuscript, the authors used ChatGPT 5.5 Thinking for language polishing, LaTeX formatting assistance, and editorial refinement of author-written text. All scientific content, methodological decisions, experiments, analyses, and final interpretations were reviewed, verified, and approved by the authors.

Separately, ChatGPT 5.5 Thinking and Gemini 3.1 Pro were used as part of the reported research methodology for LLM-guided prediction refinement and teacher-signal generation, as described in Section 4. Only de-identified radiology reports and first-stage model predictions were provided to these systems. No patient identifiers, raw MRI files, whole-slide pathology images, or private metadata were provided.

After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] D. N. Louis, A. Perry, P. Wesseling, D. J. Brat, I. A. Cree, D. Figarella-Branger, C. Hawkins, H. K. Ng, S. M. Pfister, G. Reifenberger, R. Soffietti, A. von Deimling, D. W. Ellison, The 2021 WHO Classification of Tumors of the Central Nervous System: A Summary, *Neuro-Oncology* 23 (2021) 1231–1251. doi:10.1093/neuonc/noab106.
- [2] A. Ben Abacha, J. E. Heras Rivera, D. K. Low, W.-w. Yim, J. Ruzevick, D. Child, M. Kurt, Overview of the MEDIQA-CORE 2026 Task 1 on Brain Tumor Subtype Classification, in: *CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026*.
- [3] J. E. Heras Rivera, D. K. Low, X. Xiong, J. J. Ruzevick, D. D. Child, W.-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). arXiv:2603.03618.
- [4] A. Kondepudi, A. Rao, C. Zhao, Y. Lyu, S. Harake, S. Banerjee, R. Joshi, A.-K. Meissner, R. Hou, C. Jiang, A. Chowdury, A. Srinivasan, B. Athey, V. Gulani, A. Pandey, H. Lee, T. Hollon, Health System Learning Achieves Generalist Neuroimaging Models, *arXiv Preprint arXiv:2511.18640* (2025). arXiv:2511.18640.
- [5] 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* 630 (2024) 181–188. doi:10.1038/s41586-024-07441-w.
- [6] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, A. Babiker, N. Schärli, A. Chowdhery, P. Mansfield, D. Demner-Fushman, B. Agüera y Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, V. Natarajan, Large Language Models Encode Clinical Knowledge, *Nature* 620 (2023) 172–180. doi:10.1038/s41586-023-06291-2.
- [7] L. van Santwijk, V. Kouwenberg, F. Meijer, M. Smits, D. Henssen, A Systematic Review and Meta-Analysis on the Differentiation of Glioma Grade and Mutational Status by Use of

Perfusion-Based Magnetic Resonance Imaging, *Insights into Imaging* 13 (2022) 102. doi:10.1186/s13244-022-01230-7.

- [8] H. Fan, Y. Luo, F. Gu, B. Tian, Y. Xiong, G. Wu, X. Nie, J. Yu, J. Tong, X. Liao, Artificial Intelligence-Based MRI Radiomics and Radiogenomics in Glioma, *Cancer Imaging* 24 (2024) 36. doi:10.1186/s40644-024-00682-y.
- [9] M. P. Nasrallah, J. Zhao, C. C. Tsai, D. Meredith, E. Marostica, K. L. Ligon, J. A. Golden, K.-H. Yu, Machine Learning for Cryosection Pathology Predicts the 2021 WHO Classification of Glioma, *Med* 4 (2023) 526–540.e4. doi:10.1016/j.medj.2023.06.002.
- [10] W. Wang, Y. Zhao, L. Teng, J. Yan, Y. Guo, Y. Qiu, Y. Ji, B. Yu, D. Pei, W. Duan, M. Wang, L. Wang, J. Duan, Q. Sun, S. Wang, H. Duan, C. Sun, Y. Guo, L. Luo, Z. Guo, F. Guan, Z. Wang, A. Xing, Z. Liu, H. Zhang, L. Cui, L. Zhang, G. Jiang, D. Yan, X. Liu, H. Zheng, D. Liang, W. Li, Z.-C. Li, Z. Zhang, Neuropathologist-Level Integrated Classification of Adult-Type Diffuse Gliomas Using Deep Learning From Whole-Slide Pathological Images, *Nature Communications* 14 (2023) 6359. doi:10.1038/s41467-023-41195-9.
- [11] 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 (2021) 555–570. doi:10.1038/s41551-020-00682-w.
- [12] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, HeMIS: Hetero-Modal Image Segmentation, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, volume 9901 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 469–477. doi:10.1007/978-3-319-46723-8_54.
- [13] J.-H. Choi, J.-S. Lee, EmbraceNet: A Robust Deep Learning Architecture for Multimodal Classification, *Information Fusion* 51 (2019) 259–270. doi:10.1016/j.inffus.2019.02.010.
- [14] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, G. Carneiro, Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15878–15887. doi:10.1109/CVPR52729.2023.01524.
- [15] R. Wu, H. Wang, H.-T. Chen, G. Carneiro, Deep Multimodal Learning With Missing Modality: A Survey, *Transactions on Machine Learning Research* (2026). URL: <https://openreview.net/forum?id=tc7RFcx4hT>.
- [16] H. Nori, N. King, S. M. McKinney, D. Carignan, E. Horvitz, Capabilities of GPT-4 on Medical Challenge Problems, *arXiv Preprint arXiv:2303.13375* (2023). arXiv:2303.13375.
- [17] D. Reichenpfader, H. Müller, K. Denecke, A Scoping Review of Large Language Model Based Approaches for Information Extraction From Radiology Reports, *npj Digital Medicine* 7 (2024) 222. doi:10.1038/s41746-024-01219-0.
- [18] G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network, *arXiv Preprint arXiv:1503.02531* (2015). arXiv:1503.02531, NIPS 2014 Deep Learning Workshop.
- [19] C.-Y. Hsieh, C.-L. Li, C.-k. Yeh, H. Nakhost, Y. Fujii, A. Ratner, R. Krishna, C.-Y. Lee, T. Pfister, Distilling Step-by-Step! Outperforming Larger Language Models With Less Training Data and Smaller Model Sizes, in: *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 8003–8017. doi:10.18653/v1/2023.findings-acl.507.

A. LLM Prompt Templates

We used two prompt variants for report-based prediction. The first variant used only the de-identified radiology report. The second variant used both the de-identified radiology report and the first-stage model prediction. Both prompts used the same task definitions, label mappings, reasoning instructions, and structured output schema. Subject-specific values were inserted into the placeholders shown below.

A.1. Prompt A: Report-Only LLM Prediction

You are assisting with MEDIQA-CORE Task 1: Brain Tumor Subtype Classification.

Your task is to predict three labels for each subject based only on the de-identified radiology report.

Task 1: Molecular grouping (level1_pred)

0 = GBM / IDH-wt.

This includes glioblastoma and IDH-wildtype high-grade glioma.

1 = IDH-mutant diffuse glioma.

This includes astrocytoma, IDH-mutant, and oligodendroglioma.

2 = Non-diffuse / pediatric-type tumor.

This includes midline glioma H3-altered, pilocytic astrocytoma, and MAPK-driven tumors.

Task 2: LGG vs HGG (lgghgg_pred)

0 = LGG.

Low-grade glioma, corresponding primarily to WHO grade 2.

1 = HGG.

High-grade glioma, corresponding to WHO grades 3 and 4.

Task 3: WHO grade (who_grade_pred)

0 = WHO grades 1--2.

1 = WHO grade 3.

2 = WHO grade 4.

Important reasoning instructions:

- Use medical and radiological evidence from the report.
- Do not rely only on dataset statistics or class frequency.
- Pay attention to enhancement, necrosis, edema, mass effect, infiltrative growth, multifocality, midline location, and tumor location.
- Check consistency between LGG/HGG and WHO grade.
- In general, WHO grade 0 should correspond to LGG, while WHO grade 1 or 2 should correspond to HGG.
- Be especially careful when distinguishing WHO grade 3 from WHO grade 4.
- If evidence is incomplete, choose the most plausible label and state the uncertainty.

Input:

subject_id: {SUBJECT_ID}

De-identified radiology report: {RADIOLOGY_REPORT}

Return the answer as a CSV-style table with exactly these columns:

subject_id, key_radiology_evidence, level1_pred, lgghgg_pred, who_grade_pred, confidence, short_explanation

Return only the table.

A.2. Prompt B: First-Stage Prediction + Report LLM Refinement

You are assisting with MEDIQA-CORE Task 1: Brain Tumor Subtype Classification.

Your task is to revise or confirm the first-stage model prediction using the de-identified radiology report. The first-stage model prediction is a structured prior, not ground truth. Use it as helpful evidence, but override it if the radiology report provides stronger medical evidence for another label.

Task 1: Molecular grouping (level1_pred)

0 = GBM / IDH-wt.

This includes glioblastoma and IDH-wildtype high-grade glioma.

1 = IDH-mutant diffuse glioma.

This includes astrocytoma, IDH-mutant, and oligodendroglioma.

2 = Non-diffuse / pediatric-type tumor.

This includes midline glioma H3-altered, pilocytic astrocytoma, and MAPK-driven tumors.

Task 2: LGG vs HGG (lgghgg_pred)

0 = LGG.

Low-grade glioma, corresponding primarily to WHO grade 2.

1 = HGG.

High-grade glioma, corresponding to WHO grades 3 and 4.

Task 3: WHO grade (who_grade_pred)

0 = WHO grades 1--2.

1 = WHO grade 3.

2 = WHO grade 4.

Important reasoning instructions:

- Use medical and radiological evidence from the report.
- Use the first-stage model prediction as a prior, not as the final answer.
- Do not rely only on dataset statistics or class frequency.
- Pay attention to enhancement, necrosis, edema, mass effect, infiltrative growth, multifocality, midline location, and tumor location.
- Check consistency between LGG/HGG and WHO grade.
- In general, WHO grade 0 should correspond to LGG, while WHO grade 1 or 2 should correspond to HGG.
- Be especially careful when distinguishing WHO grade 3 from WHO grade 4.
- If you change the first-stage model prediction, explain the radiological evidence that supports the change.
- If evidence is incomplete, choose the most plausible label and state the uncertainty.

Input:

subject_id: {SUBJECT_ID}

First-stage model prediction:

level1_pred: {FIRST_STAGE_LEVEL1_PRED}

lgghgg_pred: {FIRST_STAGE_LGGHGG_PRED}

who_grade_pred: {FIRST_STAGE_WHO_GRADE_PRED}

De-identified radiology report: {RADIOLOGY_REPORT}

Return the answer as a CSV-style table with exactly these columns:

subject_id, key_radiology_evidence, level1_pred, lgghgg_pred, who_grade_pred, confidence, short_explanation

Return only the table.

B. Model and Training Hyperparameters

This appendix summarizes the implementation details of the NeuroVietCH system used for the submitted runs.

B.1. Input Representations

- **MRI representation:** provided padded T1CE MRI embeddings.
- **MRI token length:** 1337 padded tokens.
- **MRI feature aggregation:** mean pooling over MRI tokens.
- **MRI feature dimension:** 768.
- **Pathology representation:** provided GigaPath slide embeddings.
- **Pathology aggregation:** mean aggregation across available slides for each subject.
- **Pathology feature dimension:** 768.
- **Missing pathology handling:** learnable missing-pathology token used when pathology embeddings were unavailable.

B.2. Model Architecture

- **MRI expert:** linear modality-specific probe trained on the mean-pooled MRI feature vector, mapping 768 input features to the task-specific number of classes.
- **Pathology expert:** linear modality-specific probe trained on the subject-level pathology feature vector, mapping 768 input features to the task-specific number of classes.
- **Expert usage:** selected MRI and pathology expert checkpoints were frozen before fusion training.
- **Fusion module:** gated logit fusion using frozen expert logits.
- **Residual pathway:** residual fusion MLP using concatenated MRI-pathology features.
- **Output heads:** task-specific classifiers for molecular grouping, LGG/HGG status, and WHO grade.
- **Probe hidden sizes:** none; both MRI and pathology experts were linear probes with zero hidden layers.
- **Gate hidden size:** 16.
- **Residual fusion hidden size:** 64.
- **Activation function:** ReLU for the gating network and GELU for the residual fusion pathway.
- **Dropout:** 0.0 for unimodal probes and 0.4 for the residual fusion pathway.
- **Normalization:** LayerNorm.

B.3. Optimization

- **Optimizer:** Adam.
- **Learning rate:** 1×10^{-3} .
- **Weight decay:** 1×10^{-4} .
- **Batch size:** 8.
- **Training schedule:** 512 epochs without early stopping.
- **Class imbalance handling:** class-balanced weighted sampler during training.
- **EMA strategy:** an exponential moving average of model weights was maintained throughout training with decay 0.999.
- **Final checkpoint:** the EMA checkpoint was used for final inference.
- **Loss function:** cross-entropy for supervised task training; teacher-student distillation loss as described in Section 4.

B.4. Checkpoint Selection

- **MRI expert checkpoint:** selected based on consistently strong train/validation Macro-F1.
- **Pathology expert checkpoint:** selected based on downstream fusion behavior, rather than only the earliest standalone validation peak.
- **Fusion checkpoint:** selected using EMA-based checkpoint selection to reduce sensitivity to short-term validation fluctuations.
- **Final Task 1 model:** trained using the selected frozen experts, class-balanced sampling, gated logit fusion, residual feature fusion, and task-specific output heads.

B.5. LLM-Guided Distillation

- **Teacher models:** ChatGPT-guided and Gemini-guided teacher variants were evaluated.
- **Best teacher:** ChatGPT-guided teacher.
- **Teacher inputs:** first-stage model predictions and de-identified radiology reports.
- **Teacher outputs:** refined task labels, soft targets when available, and concise evidence summaries.
- **Student model:** same multimodal fusion architecture as the first-stage fusion model.
- **Inference:** final distilled student does not require LLM calls.
- **Distillation weights:** $\alpha = 0.8$ and $\lambda = 0.5$.

B.6. Hardware and Software Environment

- **Hardware:** NVIDIA RTX A6000 GPU with 48 GB memory, 251 GB system RAM, and 16 CPU cores.
- **Operating system:** Ubuntu 22.04 with Linux kernel 6.5.0-41-generic.
- **Python version:** Python 3.11.12.
- **Deep learning framework:** PyTorch 2.10.0, with `torch>=2.4.0`, `lightning>=2.4.0`, and `monai==1.3.2`.
- **Main libraries:** MONAI 1.3.2, timm 1.0.26, scikit-learn 1.0.0, NumPy 2.4.3, and pandas 3.0.2.

C. Privacy and Data Use

The CoRe-BT data were used under the MEDIQA-CORE Task 1 shared-task setting for model development, validation, and official submission. All experiments were conducted using the data and derived representations provided by the organizers, including MRI embeddings, pathology embeddings, radiology reports, and task labels.

For LLM-guided refinement, the input consisted of de-identified radiology reports and first-stage model predictions. The LLMs were not provided with raw MRI volumes, whole-slide pathology images, pathology tiles, or subject-level clinical metadata beyond the information contained in the de-identified reports. The resulting LLM outputs were used as structured teacher signals for prediction refinement and distillation.

The inference code and checkpoint for the submitted system are publicly available at <https://github.com/candleMind/corebt-llm-distill>. The CoRe-BT dataset is not redistributed with this work. The repository excludes private dataset files, raw MRI volumes, whole-slide pathology images, radiology reports, pathology tiles, subject metadata, and other restricted challenge materials. Additional training scripts, ablations, and extended analyses may be released in a follow-up detailed study.