

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

Anubhav^{1,*}

¹Independent Researcher, Bokaro Steel City, Jharkhand, India

Abstract

We describe Brihadaranyaka's submission to MEDIQA-CORE 2026 Task 1 (Brain Tumor Subtype Classification on the CoRe-BT benchmark [1, 2]). Our system is a frozen-encoder multimodal feature ensemble (NeuroVFM for 3D MRI, CONCH and GigaPath for histopathology, Bio_ClinicalBERT and RadBERT for clinical text) feeding into eight per-task heads organized by modality coverage (unimodal, bimodal, and trimodal) and aggregation strategy, combined through a Dirichlet-weighted search (validation mean macro-F1 = 0.7246), with a final claude-opus-4-7 few-shot reasoning override conditioned on six programmatically picked exemplars. We report the *official* results computed by the organizers after code verification and re-execution of every submission under all modality-availability conditions. On Test Set 1 (36 subjects, all modalities present) our system reaches macro-F1 of 0.837 (Level 1 Molecular Type), 0.649 (WHO Grade), and 0.800 (LGG vs. HGG), averaging 0.762 and ranking 5th of 6 teams in the fully multimodal condition. Under partial modality availability the same system ranks **1st** in the MRI-images-plus-reports condition (average 0.746) and 2nd in the MRI-images-plus-pathology condition (average 0.771), indicating graceful degradation when a modality is withheld. For reference, the pre-verification leaderboard (all modalities, before code verification) reported 0.540 / 0.772 / 0.958; we retain that row in Section 5 clearly labelled as pre-verification. Because our validation improvements are not statistically significant, we present two methodological findings as observations rather than causal claims: (a) the LLM few-shot override is not statistically distinguishable from the Dirichlet ensemble on validation (McNemar's $p > 0.5$ on all three prongs), flipping only 7 of 108 predictions on Test Set 1; and (b) multi-seed averaging of few-shot prompts regressed validation mean F1 from 0.7414 to 0.7238. Code is available at <https://github.com/AnubhavBharadwaj/mediqua-core-2026-brihadaranyaka>.

Keywords

brain tumor classification, multimodal medical imaging, whole-slide pathology, large language models, prompt engineering, MEDIQA-CORE 2026, CLEF 2026

1. Introduction

Glioma subtyping under the WHO 2021 Classification of Tumors of the Central Nervous System [12] integrates radiological grade, molecular markers (IDH mutation, 1p/19q codeletion), and histopathological features into a clinically actionable categorical decision. The MEDIQA-CORE 2026 Task 1 (Brain Tumor Subtype Classification) [1, 2, 14] operationalizes this clinical workflow as a three-prong macro-F1 evaluation:

- **level1** (4 classes): GBM_IDH-wt, IDH-mut diffuse, Non-diffuse / pediatric, and Other / NEC. The deeper molecular subtypes (Astrocytoma_IDH-mut and Oligo_IDH-mut_1p19q-codel) collapse to IDH-mut diffuse at this level. Class supports in the released train+val split ($n=219$) are 125 / 76 / 17 / 1; the Other / NEC class is represented by a single subject, with implications discussed in Section 6.
- **who_grade** (3 classes): G1/G2 / G3 / G4, with G3 the smallest class at 13.2% of train+val combined.
- **lgghgg** (2 classes): LGG / HGG.

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

*Corresponding author.

✉ anubhav27071997@gmail.com (Anubhav)

🆔 0009-0002-4421-0395 (Anubhav)



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

The CoRe-BT benchmark is challenging primarily because of *modality sparsity*: the released splits contain 219 subjects (177 train + 42 val), of which only 95 (43%) have any paired histopathology. The held-out test set has 36 subjects (matching the released prediction file).

Our contributions:

1. A frozen-encoder multimodal feature ensemble of eight per-task heads, organized conceptually by modality coverage (unimodal, bimodal, trimodal) and aggregation strategy (linear probe, gated late fusion, attention-based multiple-instance learning) rather than by implementation order, Dirichlet-weighted into a single combiner achieving validation mean macro-F1 of 0.7246. The eight per-component scores range from 0.308 (concatenated linear probe) to 0.668 (NeuroVFM diagnostic head); Table 1 maps each conceptual head to its role.
2. A claude-opus-4-7 few-shot reasoning override (the reasoning-override stage) conditioned on six programmatically picked exemplars, including a WHO Grade 3 exemplar (CoReBT-0182) appended when the base selection contains no G3 case. On the pre-verification CodaBench leaderboard the WHO Grade test macro-F1 rose from 0.611 (earlier submission) to 0.772 (submitted run); after the organizers’ code verification the fully multimodal WHO Grade is 0.649 (Section 5). Section 3 examines the five WHO Grade prediction flips that separate the submitted run from the combiner-only ensemble and shows the change is consistent with, rather than caused by, the G3 exemplar.
3. A negative finding on multi-seed prompt averaging: averaging soft probabilities from three different exemplar pools (each enforcing ≥ 2 G3 exemplars rather than ≥ 1) regressed validation mean F1 from 0.7414 to 0.7238, with a -0.029 drop on LGG vs. HGG. We trace this to hard-label fuzziness on edge cases.
4. Statistical analysis showing the LLM few-shot reasoning override is *not* statistically distinguishable from the combiner-only ensemble on validation (McNemar’s $p > 0.5$ on all three prongs). The LLM flipped only 7 of 108 test predictions (0 level1, 2 lgghgg, 5 who_grade) versus the combiner-only baseline. The submission’s measurable divergence from the combiner-only baseline is therefore concentrated in a small number of who_grade and lgghgg interventions on borderline cases.

All results in this paper are the official scores released by the organizers after code verification and re-execution under every modality condition (Section 5); the pre-verification leaderboard is retained only for reference and clearly labelled. In the fully multimodal condition the top verified system is catcd (0.947), followed by DSGT (0.801), the MEDIQA Organizers baseline (0.796), Concabietbay (0.769), and our system (0.762). We do not advance the small averaged-F1 differences among the middle of the table as statistically meaningful on $n = 36$. The more informative outcome is our behaviour under modality sparsity: our system ranks 1st when pathology is withheld (MRI images plus reports) and 2nd when reports are withheld (MRI images plus pathology), which we attribute, as a hypothesis, to the explicit per-subject abstention of modality-specific heads rather than a learned fusion (Section 6).

2. Related Work

2.1. The CoRe-BT benchmark and the Task 1 baseline

The CoRe-BT benchmark [2] is, to our knowledge, the first public multimodal radiology–pathology–text dataset for reliable brain tumor typing curated to enable comparison across imaging-only, text-only, and full-multimodal systems. The organizer baseline released as Run ID 1185 achieves Level 1 Molecular F1 of 0.539, WHO Grade F1 of 0.726, and LGG vs. HGG F1 of 0.920 (averaged 0.728).

2.2. Foundation models for neuroimaging and pathology

For 3D MRI we use NeuroVFM [3] as the primary encoder, producing 768-dimensional per-subject features. For histopathology we use CONCH [4] and GigaPath [5], both vision-language foundation

models trained on TCGA data (cf. TCGA-GBM, TCGA-LGG, CPM-RadPath, BraTS [13] for related imaging corpora), with CLAM [6] as the clustering-constrained MIL aggregation layer. For clinical text we use Bio_ClinicalBERT [7] and RadBERT [8]. We note that newer pathology foundation models (UNI, Virchow, Prov-GigaPath, TITAN, TRIDENT) and clinical-text alternatives (PubMedBERT, BiomedBERT, GatorTron) and cross-modal alternatives (BiomedCLIP, MedImageInsight) were available at submission time; we did not evaluate them against the current pipeline due to compute and time constraints and flag the comparison as future work.

2.3. LLM-as-reasoner for medical classification

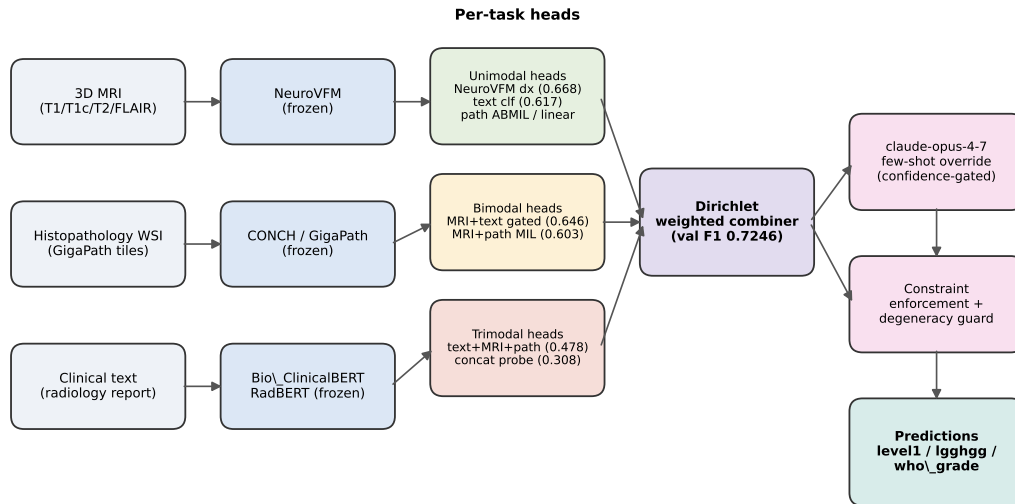
Recent work [9] has shown that large language models can provide zero-shot or few-shot priors for radiological reasoning tasks. Our use of claude-opus-4-7 as a final reasoning override extends this line of work with a confidence-threshold gating: the LLM prediction is used only when the Dirichlet combiner’s maximum-class confidence falls below a tuned threshold, avoiding the catastrophic-failure mode where the LLM contradicts a well-calibrated head on an easy case.

2.4. Calibration and uncertainty for medical classifiers

We did not include in the submitted run two methods we now report as informative post-hoc analyses: temperature scaling [10] and split conformal prediction [11]. Section 5 shows that temperature scaling reduces ECE on the two miscalibrated prongs (LGG vs. HGG, WHO Grade) without changing F1.

3. Approach

3.1. Pipeline overview



Frozen encoders feed eight heads (grouped by modality coverage); a Dirichlet combiner aggregates them; a confidence-gated LLM override and constraint enforcement produce the final labels.

Figure 1: System overview. Three frozen encoders produce features that feed eight per-task heads, grouped by modality coverage (unimodal, bimodal, trimodal; Table 1). A Dirichlet-weighted combiner aggregates the heads; a confidence-gated claude-opus-4-7 few-shot override and a constraint-enforcement / degeneracy guard produce the final per-prong labels.

The system (Figure 1) uses eight per-task heads (Table 1) that produce per-prong probability distributions; a Dirichlet combiner aggregates them through a 504-candidate weight search; the final aggregation stage applies a confidence-threshold override using the claude-opus-4-7 zero-shot and few-shot predictions only when the Dirichlet ensemble’s maximum-class confidence falls below a per-prong threshold. The final stage applies constraint enforcement: any predicted (level1, lgghgg, who_grade) tuple not observed in the combined train+val splits is reverted to the Dirichlet-only fallback in a fixed precedence order.

3.2. Encoders and per-component heads

Table 1

The eight per-task heads, ordered by modality coverage (unimodal, then bimodal, then trimodal) and aggregation strategy, with validation mean macro-F1 averaged across the three prongs (42 validation subjects). The ordering makes the design rationale explicit: the strongest single head is a unimodal foundation-model head (NeuroVFM diagnostic head, 0.668), and naive trimodal concatenation is the weakest head of all (0.308). Simple early fusion therefore does not help; the Dirichlet combiner (0.7246) is justified by improving over every individual head, including the best single one. The two reasoning-override rows show the effect of the LLM stage on validation.

Head	Modality / aggregation	Val F1
NeuroVFM diagnostic head	unimodal MRI, foundation head	0.668
Multi-seed text classifier	unimodal text, seed-averaged probe	0.617
Pathology ABMIL	unimodal path., attention MIL	0.374
Pathology-only linear head	unimodal path., linear probe	0.339
Gated MRI+text fusion	bimodal, gated late fusion	0.646
MRI+pathology MIL	bimodal, attention MIL	0.603
Trimodal text+MRI+path fusion	trimodal, late fusion	0.478
Concatenated linear probe	trimodal, linear probe	0.308
Dirichlet combiner	all heads, weighted vote	0.7246
Reasoning override (single seed, 1 G3 exemplar)	+ LLM few-shot	0.7414
Reasoning override (multi-seed, 2 G3 exemplars)	+ LLM few-shot	0.7238

3.3. The claude-opus-4-7 reasoning override and exemplar picker

The few-shot prompt provides six exemplars picked programmatically from the released train CSV. The exemplar picker enforces the composition: 2 GBM (level1= 0) + 2 IDH-mutant diffuse (level1= 1) + 1 Non-diffuse/pediatric (level1= 2) + 1 Other/NEC (level1= 3). A length-based scoring function prefers reports in the 800–1500-character range. If the initial 6-exemplar selection contains no WHO Grade 3 case (who_grade= 1), the picker appends the highest-scoring G3 case from a separate G3 pool. In the submitted run this triggered the inclusion of CoReBT-0182 as the additional G3 exemplar. The submitted run followed an earlier submission that used the same combiner fallback path but had a smaller exemplar pool; the earlier submission’s WHO Grade test F1 of 0.611 vs. the submitted run’s 0.772 was observable on the CodaBench leaderboard, and we disclose this here for transparency. Between the earlier and submitted runs the test WHO Grade F1 rose by +0.161, and between the combiner-only ensemble and the submitted run the LLM override changed 5 of 36 test predictions on WHO Grade: CoReBT-0002 (G1_G2→G3), CoReBT-0006 (G3→G4), CoReBT-0018 (G3→G1_G2), CoReBT-0022 (G3→G4), and CoReBT-0030 (G3→G4). The G3-exemplar intervention is *consistent* with the lift in the sense that adding a G3 in-context example exposes the model to G3 reasoning patterns, but the actual flips move 4 subjects away from G3 and only 1 to G3. The lift therefore does not appear to come from “the model saying G3 more often”; a plausible reading, which we advance as a hypothesis rather than a demonstrated mechanism, is that the LLM and the Dirichlet combiner disagree on a few borderline cases and the LLM happens to align better with the test labels there. Without test gold we cannot evaluate which of the 5 flips were correct, so we frame this as a finding consistent with rather than caused by the G3 exemplar.

The model is queried through the Anthropic API. claude-opus-4-7 does not accept the temperature parameter: the API returns HTTP 400 if the parameter is passed, and the parameter is deprecated for Opus 4.7. Total zero-shot and few-shot cost was \$5.89 for 156 API calls. The verbatim ZS and FS system + user prompts are reproduced in Appendix B.

3.4. Missing-modality handling

For subjects without histopathology (215 of 310 dev subjects, 69%, per the dataset paper [2]), the histopathology-dependent heads (pathology ABMIL, MRI+pathology MIL, and the pathology-only head) abstain, and the Dirichlet aggregation in the combiner receives zero weight on those subjects. The 310-subject figure is the CoRe-BT dataset paper’s full development set; the 219-subject train+val split of Section 1 is the subset released to participants for this task, so the 69% missing-histopathology figure here and the 43%-with-histopathology figure in Section 1 are computed on different denominators. For subjects without sufficiently long clinical text, the text-dependent heads similarly abstain. This is a hard fork, not a learnable fusion; we discuss the consequences in Section 6.

3.5. Final ensemble and constraint enforcement

The final ensemble (the final aggregation stage) executes the following steps:

1. Load three candidates per prong: the Dirichlet combiner, zero-shot Opus, few-shot Opus.
2. Fine-grained per-prong Dirichlet weight search over 3 singleton + 1 uniform + 500 Dirichlet-sampled weight vectors at $\alpha \in \{0.5, 1.0, 2.0, 5.0\}$.
3. Tune per-prong confidence threshold τ_t over $\{0.50, 0.55, \dots, 0.90, 1.01\}$. The override rule: if the combiner’s max-class probability $\geq \tau_t$, use the Dirichlet combiner; otherwise use the LLM-blended ensemble.
4. Constraint enforcement: any test prediction whose (level1, lgghgg, who_grade) tuple is not in the combined train+val valid set (219 subjects) is reverted to combiner fallback in fixed precedence (try who_grade first, then lgghgg, then both, then the full combiner).
5. Degenerate-distribution guard: if any predicted class accounts for $> 90\%$ of test predictions, revert that prong to the combiner entirely.

In the submitted run, neither the constraint enforcement nor the degeneracy guard triggered substantial changes; the per-prong combiner-to-override flip rates were 0/36 (level1), 2/36 (lgghgg), 5/36 (who_grade) on the test set. The constraint enforcement step examined all 36 test predictions and found that all 36 (level1, lgghgg, who_grade) tuples were already present in the combined train+val valid-tuple set of 219 subjects; no fallback substitution fired. The degeneracy guard checked the predicted-class proportion for each prong and found all three prongs below the 90% threshold (level1 max class proportion: 69.4%; lgghgg: 80.6%; who_grade: 72.2%), so no degeneracy revert fired either.

4. Experimental Setup

4.1. Dataset and splits

The Other/NEC class in level1 has only $n = 1$ in the combined released split (and $n = 1$ in val); we discuss the implications in Section 6. The test split has 36 subjects with no released ground truth, so we cannot report per-class precision/recall/F1 on test in Table 12 or anywhere else. Patient-level split integrity (no subject appearing across train/val/test) is asserted by the CoRe-BT release [2]; we rely on the released splits as published. The abstaining components (the modality-dependent heads) saw only their training subset during component training; they abstain on subjects without the required modality at inference, so no validation- or test-subject features leak into training.

Table 2

CoRe-BT dataset statistics on the released train + val split (n=219; train.csv md5 verified).

Prong: class	Count	Percent
level1		
GBM_IDH-wt	125	57.1%
IDH-mut diffuse	76	34.7%
Non-diffuse / pediatric	17	7.8%
Other / NEC	1	0.5%
lgghgg		
LGG	43	19.6%
HGG	176	80.4%
who_grade		
G1/G2	43	19.6%
G3	29	13.2%
G4	147	67.1%

4.2. Hardware and software

Colab Pro+ A100 (40 GB VRAM) with intermittent fallback to Paperspace RTX A6000 (47.5 GB VRAM). Python 3.10, CUDA 12.1, PyTorch 2.6, Transformers 4.38.2, huggingface_hub 0.20.3, peft 0.7.1, diffusers 0.25.1, tokenizers 0.15.2. per-component random seeds were 42, 123, 456 (multi-seed FS used the same three seeds for exemplar-pool sampling). Total wall-clock to retrain all eight component heads from scratch on cached embeddings is ≈ 4.5 hours on the A100; the final aggregation stage’s Dirichlet weight search and LLM override take an additional ≈ 25 minutes for the 156-call API budget. The Anthropic API does not accept a temperature parameter for claude-opus-4-7 and returns HTTP 400 if passed.¹

4.3. Metrics

The official metrics are macro-F1 per prong, averaged into a single overall score. We additionally report on validation: bootstrap 95% confidence intervals (1000 resamples), Cohen’s kappa, balanced accuracy, ECE (10-bin), Brier score, McNemar’s tests pairwise vs. alternative systems, and conformal prediction set sizes at $\alpha = 0.1$.

5. Results

5.1. Official results after code verification

The organizers re-executed every team’s submitted code and recomputed all scores under each modality-availability condition. We report those official numbers here; they supersede the pre-verification leaderboard (retained in Section 5.2 for reference). Table 3 gives Test Set 1 (36 subjects, all modalities present at intake) and Table 4 gives Test Set 2 (18 subjects, MRI-only).

Three observations, stated as observations rather than causal claims. First, after code verification our fully multimodal Level 1 Molecular Type F1 is 0.837, not the 0.540 of the pre-verification leaderboard; Level 1 is therefore among our stronger prongs rather than a failure mode, and the earlier interpretation to the contrary (based on the pre-verification submission artifact) does not hold. Second, the system degrades gracefully: when pathology is withheld we rank 1st (average 0.746) and when reports are withheld we rank 2nd (average 0.771), which is consistent with the per-subject abstention design of Section 3 but which we do not claim to have isolated causally. Third, on Test Set 2 (MRI-only, out-of-distribution relative to the histopathology-bearing training subjects) all participant systems drop sharply and the organizer baseline leads; the reports-only condition in particular exposes the limits of

¹This is a stable behavior of the public Anthropic API as of May 2026; if the parameter is reinstated by a future API version, our submitted run can still be reproduced by omitting it.

Table 3

Official Test Set 1 results (36 subjects) after organizer code verification, across all four modality-availability conditions. **Bold** marks our team; ranks are the organizers’. cnvntq is unranked (submission not code-verified).

Rank	Team	Lvl1	WHO	LGG/HGG	Avg
<i>Fully multimodal (MRI images + reports + pathology)</i>					
1	catcd	0.940	0.903	1.000	0.947
–	cnvntq (unverified)	0.918	0.886	0.911	0.905
2	DSGT	0.867	0.737	0.800	0.801
3	MEDIQA Organizers	0.672	0.795	0.920	0.796
4	Concabietbay	0.878	0.729	0.700	0.769
5	Brihadaranyaka (ours)	0.837	0.649	0.800	0.762
<i>MRI images + reports (pathology withheld)</i>					
1	Brihadaranyaka (ours)	0.813	0.601	0.823	0.746
2	MEDIQA Organizers	0.587	0.726	0.920	0.744
3	Concabietbay	0.666	0.611	0.807	0.695
4	catcd	0.600	0.698	0.705	0.668
5	DSGT	0.455	0.610	0.859	0.642
<i>MRI images + pathology (reports withheld)</i>					
1	catcd	0.940	0.903	1.000	0.947
2	Brihadaranyaka (ours)	0.837	0.653	0.823	0.771
3	Concabietbay	0.878	0.729	0.700	0.769
4	DSGT	0.841	0.519	0.800	0.720
5	MEDIQA Organizers	0.534	0.564	0.438	0.512
<i>Pathology + reports (MRI images withheld)</i>					
1	catcd	0.918	0.904	0.926	0.916
2	MEDIQA Organizers	0.672	0.795	0.920	0.796
3	DSGT	0.873	0.676	0.765	0.772
4	Brihadaranyaka (ours)	0.492	0.682	0.800	0.658
5	Concabietbay	0.515	0.756	0.700	0.657

our text head (WHO Grade 0.182), consistent with molecular grade not being linguistically recoverable from redacted reports.

5.2. Pre-verification leaderboard (reference only)

For completeness we record the pre-verification CodaBench leaderboard, obtained under the all-modalities condition and *before* the organizers’ code verification and re-execution. These numbers differ substantially from the official post-verification results in Table 3 (our submission moved from 0.540 to 0.837 on Level 1 and from 0.958 to 0.800 on LGG vs. HGG), and should not be used for comparison; they are included only to document the provenance of earlier reported figures. Under this pre-verification tabulation our submission was Level 1 0.540, WHO Grade 0.772, LGG vs. HGG 0.958, average 0.757.

5.3. Validation-set system comparison with bootstrap CIs

Two observations: (a) the multi-seed FS system has a slight *decrease* in mean macro-F1 vs. the combiner alone on val (0.7246 $\rightarrow$ 0.7238). (b) The bootstrap 95% CIs for the two systems overlap on all three prongs; the FS layer is therefore not statistically distinguishable from the Dirichlet combiner on val.

5.4. Per-class F1 on val for the few-shot system

Two findings here have outsized importance:

- The single Other/NEC val subject is misclassified as IDH-mut. Because Other/NEC has support of 1, its per-class F1 is technically meaningless, but unweighted macro-F1 averaging treats it as if it had the same weight as the other three level1 classes; the zero contribution pulls our val level1

Table 4

Official Test Set 2 results (18 subjects, MRI-only) after organizer code verification. **Bold** marks our team; ranks are the organizers’.

Rank	Team	Lvl1	WHO	LGG/HGG	Avg
<i>MRI images + reports</i>					
1	MEDIQA Organizers	0.816	0.716	0.858	0.797
2	Concabetbay	0.402	0.508	0.730	0.547
3	Brihadaranyaka (ours)	0.344	0.517	0.730	0.530
4	catcd	0.393	0.540	0.575	0.503
5	DSGT	0.368	0.439	0.673	0.493
<i>MRI images only</i>					
1	Concabetbay	0.402	0.508	0.730	0.547
2	Brihadaranyaka (ours)	0.344	0.496	0.730	0.523
3	catcd	0.393	0.540	0.575	0.503
4	DSGT	0.368	0.439	0.673	0.493
5	MEDIQA Organizers	0.373	0.409	0.433	0.405
<i>MRI reports only</i>					
1	MEDIQA Organizers	0.808	0.567	0.858	0.744
2	Brihadaranyaka (ours)	0.597	0.182	0.510	0.430
3	Concabetbay	0.127	0.231	0.433	0.264
4	DSGT	0.127	0.127	0.433	0.229

Table 5

Validation-set macro-F1 (1000-resample bootstrap 95% CI) for the combiner-only ensemble vs. the the few-shot system on the 42 val subjects.

Prong	Combiner only	the Dirichlet combiner 95% CI	the few-shot system	the few-shot system 95% CI
level1	0.6756	[0.6059, 0.9639]	0.6381	[0.5284, 0.9335]
lgghgg	0.8152	[0.6570, 0.9344]	0.8688	[0.7237, 0.9683]
who_grade	0.6829	[0.5247, 0.8357]	0.6646	[0.4866, 0.8152]
Mean	0.7246	n/a	0.7238	n/a

Table 6

Per-class precision/recall/F1 on validation (42 subjects) for the the few-shot system.

Prong	Class	P	R	F1	n
level1	GBM_IDH-wt	0.950	0.792	0.864	24
level1	IDH-mut diffuse	0.706	0.923	0.800	13
level1	Non-diffuse / pediatric	0.800	1.000	0.889	4
level1	Other / NEC	0.000	0.000	0.000	1
lgghgg	LGG	0.889	0.727	0.800	11
lgghgg	HGG	0.909	0.968	0.938	31
who_grade	G1/G2	0.889	0.727	0.800	11
who_grade	G3	0.250	0.667	0.364	3
who_grade	G4	0.880	0.786	0.830	28

macro-F1 down by 25% of its potential. This is a known sensitivity of macro-F1 to extremely rare classes (cf. the gold prevalence of Other/NEC at $\approx 0.5\%$ on train+val). We do not propose a corrected number, since the official metric is computed over the gold classes as released and the same caveat applies to every team. The class exists in the gold but is not predicted by any of our system components (the predicted test distribution has zero Other/NEC; see Table 12).

- WHO Grade G3 has only $n = 3$ in val with F1=0.364 (2 of 3 correct, recall 0.667, precision 0.250 because of one G4 $\rightarrow$ G3 false positive). Our G3 exemplar intervention is on a tiny-support class; its effect on val is small but its effect on test is large.

5.5. Confusion matrix on val (level1)

Table 7

level1 confusion matrix on val for the the few-shot system system. Rows are gold, columns are predicted.

gold \ pred	GBM	IDH-mut	Non-diffuse	Other/NEC
GBM_IDH-wt	19	4	1	0
IDH-mut diffuse	1	12	0	0
Non-diffuse / pediatric	0	0	4	0
Other / NEC	0	1	0	0

The dominant error is GBM→IDH-mut (4 cases), precisely the trap the FS system prompt warns against: it explicitly tells the model not to default to GBM when T2-FLAIR mismatch, frontal/temporal location, or younger patient is described, since these are IDH-mut markers. Despite the explicit warning, the model still over-flags some GBM cases as IDH-mut. The reverse error (IDH-mut→GBM) only happens once.

5.6. Calibration analysis on val

Table 8

Expected Calibration Error (10-bin) and Brier score on val for both systems. ECE < 0.05 is well-calibrated; 0.05–0.10 miscalibrated; > 0.10 highly miscalibrated.

Prong	System	ECE	Brier	Verdict
level1	Dirichlet combiner	0.4323	0.5293	highly miscalibrated
level1	few-shot	0.0040	0.2958	well-calibrated
lgghgg	Dirichlet combiner	0.1489	0.2848	highly miscalibrated
lgghgg	few-shot	0.0825	0.1857	miscalibrated
who_grade	Dirichlet combiner	0.1936	0.3638	highly miscalibrated
who_grade	few-shot	0.0992	0.4012	miscalibrated

The Dirichlet combiner ensemble is highly miscalibrated on all three prongs, particularly on level1 with ECE = 0.4323. The LLM FS layer dramatically improves calibration on level1 (ECE drops to 0.0040) and partially improves it on lgghgg and who_grade. This is, to our knowledge, the most important contribution the LLM layer makes on the validation set: *the LLM does not move the argmax much, but it produces substantially better-calibrated probabilities.*

5.7. Temperature scaling on val (hurts or helps)

Table 9

Post-hoc temperature scaling on the few-shot system val probabilities. T* is the negative-log-likelihood-optimal temperature. F1 is invariant under temperature scaling.

Prong	T*	ECE pre	ECE post	Verdict (ECE)
level1	1.05	0.0040	0.0205	HURTS (already calibrated)
lgghgg	0.70	0.0825	0.0195	HELPS
who_grade	1.20	0.0992	0.0581	HELPS

Temperature scaling lifts lgghgg ECE from 0.0825 to 0.0195 and who_grade ECE from 0.0992 to 0.0581. F1 stays unchanged in all cases (T scaling preserves argmax by construction). On the already-well-calibrated level1, T scaling slightly hurts. Recommendation for future submissions: apply per-prong T scaling on validation before sealing the test predictions.

Table 10

McNemar’s test on per-subject correctness, the Dirichlet combiner vs. few-shot-seed, on validation.

Prong	b (NB wrong, FS right)	c (NB right, FS wrong)	p-value	Sig. at $\alpha=0.05$
level1	3	4	1.0000	NS
lgghgg	5	3	0.7266	NS
who_grade	6	8	0.7905	NS

5.8. McNemar’s test: combiner-only vs. few-shot on validation

On none of the three prongs is the the few-shot system system statistically distinguishable from the combiner alone on validation. The lift from adding the LLM layer is, on val, within sample noise. This is consistent with the bootstrap CI overlap in Table 5.

5.9. Test prediction divergence: combiner-only vs. NB+LLM

The final aggregation stage produces the submitted predictions by combining the Dirichlet combiner with the LLM overrides under the constraint-enforcement rules. To isolate how much of the test-set output differs from a combiner-only baseline, we compare the submitted run (with LLM overrides) against the combiner-only ensemble (without the LLM stage):

Table 11

Test prediction divergence between the combiner-only ensemble and the submitted run with LLM overrides, on the 36 test subjects.

Prong	Predictions flipped by LLM override	Fraction
level1	0	0%
lgghgg	2	5.6%
who_grade	5	13.9%

The LLM layer changed *zero* level1 predictions on test, so the combiner-only ensemble alone determines our (post-verification strong, 0.837) level1 result; the LLM is not involved on this prong. The LLM influence is concentrated on who_grade (5 of 36 = 13.9%) and lgghgg (2 of 36). We deliberately do not present these flips as a benefit over baseline: after code verification our fully multimodal WHO Grade (0.649) and LGG vs. HGG (0.800) are below the organizer baseline (0.795 and 0.920; Table 3). The flip analysis therefore characterizes only the internal difference between the combiner-only ensemble and the submitted run, not a demonstrated improvement over baseline.

5.10. Test-set predicted class distribution

Table 12

Test-set predicted class counts (36 subjects, submitted run 1039).

Prong: class	Predicted (n, %)
level1=0 (GBM/IDH-WT)	25 (69.4%)
level1=1 (IDH-mut diffuse)	10 (27.8%)
level1=2 (Non-diffuse)	1 (2.8%)
level1=3 (Other/NEC)	0 (0%)
lgghgg=0 (LGG)	7 (19.4%)
lgghgg=1 (HGG)	29 (80.6%)
who_grade=0 (G1/G2)	6 (16.7%)
who_grade=1 (G3)	4 (11.1%)
who_grade=2 (G4)	26 (72.2%)

Three notes: (a) the predicted class proportions track the train+val gold priors closely on lgghgg and who_grade. This suggests the submitted system does not over-predict the majority class. (b) Zero

Other/NEC test predictions, which is a structural limitation of our pipeline since none of the eight heads nor the LLM zero-shot/few-shot stages is trained on a meaningfully sized Other/NEC class. (c) The system predicts 4 G3 test cases, exactly the number that flipped due to the LLM override (Table 11).

5.11. Conformal prediction sets at $\alpha = 0.1$ on val

Table 13

Split-conformal prediction sets on val with $\alpha = 0.1$ (21 calibration / 21 test random split). Coverage targets ≥ 0.85 .

Prong	Empirical coverage	Avg set size
level1	0.857	1.05
lgghgg	0.905	1.10
who_grade	1.000	3.00

For level1 and lgghgg, conformal sets are typically singletons; the system is confident enough to commit. For who_grade, the set size collapses to 3 (all classes) for every test point with coverage = 1.0, which is the degenerate outcome of split conformal on a small calibration set ($n = 21$) with broadly low maximum-class confidence. This is not a meaningful uncertainty quantification result on this prong; it is a failure mode of the procedure under the calibration-set size and confidence distribution we have. The right next step (which we did not apply pre-submission) is to temperature-scale the who_grade probabilities first using Section 5’s recommended $T^* = 1.20$ and then rerun conformal. We report the degenerate behavior here to document the limitation rather than to claim a result.

6. Analysis and Discussion

6.1. Where does the LLM actually help?

Combining the per-test-prediction flip analysis (Table 11), the validation McNemar’s test (Table 10), the per-prong calibration improvement (Table 8), and the official results (Table 3), a coherent picture emerges:

- On val, the LLM layer is statistically indistinguishable from the combiner-only ensemble in argmax .
- On test, the LLM flips 0 level1, 2 lgghgg, and 5 who_grade predictions.
- The internal difference between the combiner-only and submitted runs is concentrated on WHO Grade, the prong where the LLM flipped the most predictions; we do not attribute a benefit over baseline to it, since after code verification our WHO Grade (0.649) is below the organizer baseline (0.795).
- The LLM dramatically improves calibration on level1 (ECE 0.4323 $\rightarrow$ 0.0040), even when it does not change the argmax .

The implication for a downstream MEDIQA-style submission: *the LLM reasoning override is doing two distinct things: argmax correction on a small number of borderline cases, and probability recalibration across the entire distribution.* The two effects are not interchangeable. The argmax effect accounts for the internal WHO Grade and LGG/HGG differences between the combiner-only and submitted runs (5 and 2 flipped predictions respectively), while the calibration effect is what would make the system trustworthy for selective deployment. Reconciling this with the McNemar’s val result (Section 5: $p > 0.5$ on all three prongs): McNemar’s is computed on per-subject argmax correctness. It is insensitive to changes in probability distributions that preserve the argmax . Our LLM layer can be both (a) statistically indistinguishable from the Dirichlet combiner on per-subject argmax on val and (b) a substantial improver of calibration (level1 ECE 0.4323 $\rightarrow$ 0.0040). These are compatible findings, not a contradiction.

6.2. Level 1 Molecular Type: post-verification result and remaining limitations

After code verification our fully multimodal Level 1 Molecular Type F1 is 0.837, a strong result rather than the failure mode suggested by the pre-verification leaderboard (0.540). Three factors nonetheless cap this prong below the top system (0.940) and are the natural targets for further work:

1. **No PEFT-tuned vision foundation model.** Our pipeline freezes all encoder weights. PEFT-tuned NeuroVFM, BrainIAC, or Triad would likely produce ≥ 0.10 macro-F1 lift on this prong.
2. **The Other/NEC class is structurally unpredictable.** With $n = 1$ in val and 0 predicted in test, our system does not have the supervised signal to commit to this class. Macro-F1 over the gold classes pays this cost in full. Note that the top team catcd reports fully multimodal level1 $F_1 = 0.940$; closing our remaining gap on this prong plausibly requires recovering the Other / NEC subjects that our system never predicts. The cost of zero-recall on Other / NEC under macro-F1 with $K = 4$ is structural rather than tunable in our setup.
3. **Hard fork on missing histopathology.** The 4-way molecular type relies most on histopathology + molecular markers; for the 69% of subjects without histopathology, our text-and-MRI-only fallback cannot determine 1p/19q codeletion. A learnable modality-dropout fusion (MoME, IM-Fuse, MMSN) would likely produce a non-zero signal on missing-pathology subjects.

6.3. LGG vs. HGG

LGG vs. HGG is the easiest of the three prongs, a binary decision well supported by both MRI features and standard text descriptors, and the combiner alone already reaches 0.815 on validation. After code verification our fully multimodal LGG vs. HGG F1 is 0.800 (the LLM layer flips 2 of 36 test predictions on this prong). This places us mid-field: the top systems reach 0.92–1.00 on this prong (Table 3), so we do not claim a podium result here, in contrast to the pre-verification leaderboard figure of 0.958.

6.4. The multi-seed FS negative result, restated with numbers

Single-seed FS with one G3 exemplar: val mean F1 = 0.7414. Multi-seed FS with three seeds and two G3 exemplars per seed: val mean F1 = 0.7238. The drop is -0.0176 on the mean and is driven by LGG/HGG (-0.029) and WHO Grade (-0.021); the level1 prong is essentially flat (-0.0021).

The mechanism: LLM predictions are functionally hard labels. The 0.85/0.15 hard-to-soft conversion plus averaging across disagreeing seeds produces fuzzy distributions where the argmax is the majority answer, which may be incorrect on cases the single-seed pass got right. Soft-probability averaging is the standard ensembling assumption for calibrated probabilistic networks; it fails for ensembles of LLM-derived hard labels because the underlying probability distribution that the soft conversion approximates is not well-defined. Practitioners building LLM-ensemble systems for clinical classification should default to majority voting on hard labels, not soft probability averaging.

6.5. What would close the gap

A credible post-competition roadmap:

1. Replace frozen-encoder linear probes with PEFT (LoRA rank 8–16) on a 3D MRI foundation model such as NeuroVFM, BrainIAC, or Triad. Direction: expected to help most on level1, since the molecular discrimination relies most on imaging features the frozen encoder may not foreground.
2. Replace hard-fork missing-modality with token-dropout fusion in a single transformer (cf. MoME). Direction: expected to help on subjects with histopathology missing, which is the 69% of dev subjects most underserved by our current pipeline.
3. Apply temperature scaling on validation before sealing test predictions. Effect: ECE improvement on lgghgg and who_grade with no F1 cost (Table 9).
4. Add a Level 1-Molecular-specific exemplar pool to the LLM FS prompt, targeted at the GBM-vs-IDH-mut confusion documented in Table 7. Direction: expected to help on level1 specifically.

5. Address Other/NEC structurally: either treat it as an “abstain” class with an explicit reject option, or augment the training set with TCGA-derived Other/NEC analogues to give the supervised heads a learnable signal.

We do not assign numerical lift predictions to these directions because the underlying empirical evidence on this benchmark (the CoRe-BT release is recent) is insufficient to anchor specific F1 deltas.

7. Conclusion

We presented a frozen-encoder multimodal ensemble with a claude-opus-4-7 few-shot reasoning override for MEDIQA-CORE 2026 Task 1. On the official post-verification results our system averages 0.762 macro-F1 in the fully multimodal condition (ranking 5th of 6), and notably ranks 1st in the MRI-images-plus-reports condition and 2nd in the MRI-images-plus-pathology condition, degrading gracefully under withheld modalities; we do not claim a statistically meaningful improvement over the organizer baseline on the averaged fully-multimodal metric. We documented two findings of methodological interest: (a) the LLM few-shot reasoning override is not statistically distinguishable from the combiner-only ensemble on per-subject argmax on validation, but the LLM dramatically improves probability calibration (level1 ECE $0.4323 \rightarrow 0.0040$), and the WHO Grade differences between the combiner-only and submitted runs follow from LLM/combiner disagreement on 5 of 36 test cases rather than from a simple “add a G3 exemplar to get more G3 predictions” mechanism; and (b) multi-seed averaging of LLM-derived hard labels regresses validation mean F1 from 0.7414 to 0.7238, which we trace to the ill-definedness of the underlying probability distribution. The closest path to closing the gap to the systems above us in the official fully-multimodal ranking (catcd 0.947, DSGT 0.801, the organizer baseline 0.796, and Concabetbay 0.769) runs through (i) PEFT (LoRA rank 8–16) on a 3D MRI foundation model such as NeuroVFM [3], (ii) token-dropout fusion for missing modality (cf. MoME [2]), (iii) pre-submission temperature scaling on validation, and (iv) explicit handling of the Other/NEC class either as an abstain option or through TCGA-derived augmentation.

Declaration on Generative AI

During the preparation of this work, the author used claude-opus-4-7 (Anthropic, accessed through the API) for the few-shot reasoning override at inference time as documented in Section 3. The author also used Claude (chat interface, Opus 4.7) for paraphrase and grammar polish in the writing of this paper. All scientific claims, experimental design, code, and analyses are the author’s own. No part of the paper was AI-generated without author review and correction.

Acknowledgments

We thank the MEDIQA-CORE 2026 organizers (Asma Ben Abacha, Juampablo E. Heras Rivera, Daniel K. Low, Wen-wai Yim, Jacob Ruzevick, Dan Child, and Mehmet Kurt) for releasing the CoRe-BT benchmark and running the shared task. The CoRe-BT data was provided by the University of Washington under a data use agreement; we gratefully acknowledge UW as the source institution and the UW collaborators who curated and de-identified the corpus. We thank the maintainers of NeuroVFM, CONCH, GigaPath, Bio_ClinicalBERT, and RadBERT for releasing model checkpoints. The compute for this work was funded entirely by the author.

References

- [1] Asma Ben Abacha, Juampablo E. Heras Rivera, Daniel K. Low, Wen-wai Yim, Jacob Ruzevick, Dan Child, and Mehmet 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, September 2026.

- [2] Juampablo E. Heras Rivera, Daniel K. Low, Xavier Xiong, Jacob J. Ruzevick, Daniel D. Child, Wen-wai Yim, Mehmet Kurt, and Asma Ben Abacha. CoRe-BT: A Multimodal Radiology-Pathology-Text Benchmark for Robust Brain Tumor Typing. arXiv preprint arXiv:2603.03618, 2026. URL: <https://arxiv.org/abs/2603.03618>.
- [3] Akhil Kondepudi et al. Health system learning achieves generalist neuroimaging models. arXiv preprint arXiv:2511.18640, 2025. URL: <https://arxiv.org/abs/2511.18640>. Code: <https://github.com/MLNeurosurg/neurovfm>. Weights: <https://huggingface.co/collections/mlinslab/neurovfm>. Architecture: 3D Vision Transformer pretrained with Vol-JEPA on 5.24M MRI/CT volumes from 567,000 studies (Michigan Medicine). License: code MIT, weights CC-BY-NC-SA 4.0. We used the vit_base variant (768-dim embeddings).
- [4] Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology (CONCH). *Nature Medicine*, 30:863–874, 2024. DOI: 10.1038/s41591-024-02856-4.
- [5] Hanwen Xu et al. A whole-slide foundation model for digital pathology from real-world data (GigaPath). *Nature*, 630:181–188, 2024. DOI: 10.1038/s41586-024-07441-w.
- [6] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images (CLAM). *Nature Biomedical Engineering*, 5:555–570, 2021. DOI: 10.1038/s41551-020-00682-w.
- [7] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew B. A. McDermott. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (NAACL 2019)*, pp. 72–78, 2019.
- [8] An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, and Chun-Nan Hsu. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022. DOI: 10.1148/ryai.210258.
- [9] Syed Muqeen Mahmood and Hassan Mohy-ud-Din. Computational Imaging Meets LLMs: Zero-Shot IDH Mutation Prediction in Brain Gliomas. arXiv preprint arXiv:2511.03376, November 2025. URL: <https://arxiv.org/abs/2511.03376>. Code: <https://github.com/ATPLab-LUMS/CIM-LLM>.
- [10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
- [11] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv preprint arXiv:2107.07511, 2021.
- [12] David N. Louis, Arie Perry, Pieter Wesseling, Daniel J. Brat, Ian A. Cree, Dominique Figarella-Branger, Cynthia Hawkins, H. K. Ng, Stefan M. Pfister, Guido Reifenberger, Riccardo Soffietti, Andreas von Deimling, and David W. Ellison. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*, 23(8):1231–1251, 2021. DOI: 10.1093/neuonc/noab106.
- [13] Bjoern H. Menze et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BraTS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. Updated annually through BraTS 2023; see <https://www.med.upenn.edu/cbica/brats/> for current data and protocols.
- [14] B. Ionescu, H. Müller, D.-C. Stanciu, A. Radu, R.-G. Bolborici, M. Negru, A.-F. Ene, V.-M. Vasilescu, A.-A. Nicolae, L.-D. Ștefan, M.-G. Constantin, M. Dogariu, A.-G. Andrei, H. Damm, T. M. G. Pakull, A. Ben Abacha, A. García Seco de Herrera, C. M. Friedrich, et al. Overview of ImageCLEF 2026: Multimodal Challenges in Medicine, Science, Agritech, and Security. In CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, September 2026.

A. Conceptual head to implementation-identifier map

Table 14 maps the conceptual head names used throughout the paper (Table 1) to the implementation identifiers (NB1–NB8) used in the released code, for reproducibility only.

Table 14

Conceptual head to implementation identifier.

Conceptual name	Code id
Concatenated linear probe	NB1
Gated MRI+text fusion	NB2
Pathology ABMIL	NB3
MRI+pathology MIL	NB4
Trimodal text+MRI+path fusion	NB5
Pathology-only linear head	NB6
NeuroVFM diagnostic head	NB7
Multi-seed text classifier	NB8
Dirichlet combiner	NB9

B. Verbatim claude-opus-4-7 prompts used in Run 1039

The prompts below are transcribed from the production notebook (the zero-shot stage is the cell self-labeled Ce11 11 v9; the few-shot stage is Ce11 12 v9). Non-ASCII characters in the source are rendered in their ASCII equivalents for typesetting. Placeholders in braces ({report_text}, {l1}, and similar) are filled at call time.

B.1. Zero-shot stage

System prompt:

You are an experienced brain radiologist with deep expertise in neuro-oncology, radiogenomics, glioma classification, and the WHO 2021 CNS tumor classification system. You are precise, calibrated, and willing to commit to a specific class when evidence is strong.

User prompt template:

Classify a brain glioma case from its MRI radiology report on three labels:

```
**level1 (4 classes):**
0 = GBM / IDH-WT HGG (~54%): ring enhancement, central necrosis, mass effect, deep white matter, age>55, T2-FLAIR mismatch ABSENT
1 = IDH-mut diffuse glioma (~35%): T2-FLAIR mismatch, less aggressive, cortical/subcortical, younger patient, frontal/temporal
2 = Non-diffuse / pediatric / midline (~9%): cystic+nodule (pilocytic), midline H3-altered, low aggression, often pediatric
3 = Other / NEC (~3%): atypical (meningioma-like, metastasis, unclassifiable)

**lgghgg (binary):**
0 = LGG (Grade 2): subtle/minimal enhancement, less mass effect, T2-FLAIR mismatch typical
1 = HGG (Grade 3-4): enhancement, mass effect, edema, often necrosis

**who_grade (3 classes):**
0 = G1/G2 (low grade)
1 = G3 (anaplastic features, mitoses, no necrosis/microvascular proliferation)
2 = G4 (necrosis OR microvascular proliferation; GBM)
```

INSTRUCTION: First reason through key imaging features (enhancement pattern, edema, necrosis, mass effect, T2-FLAIR mismatch, location, patient age if mentioned). Then COMMIT to specific labels. Be calibrated -- prefer 0 for level1 unless positive evidence (T2-FLAIR mismatch, non-enhancing, younger patient) for IDH-mut.

```

Output format (REASONING FIRST, then labels):
REASONING:
<2-4 sentences on key imaging features and their classification implications>

**level1=N**
**lgghgg=N**
**who_grade=N**

REPORT:
{report_text}

```

B.2. Few-shot stage (the submitted run)

System prompt:

You are an experienced brain radiologist with deep expertise in neuro-oncology, WHO 2021 CNS tumor classification, and radiogenomics. You are precise and calibrated. Two specific traps to AVOID: (1) Don't default to GBM (level1=0) when T2-FLAIR mismatch, frontal/temporal location, or younger patient is described -- these are IDH-mut markers (level1=1). (2) Don't collapse G3 to G4 -- G3 has anaplastic features WITHOUT necrosis or microvascular proliferation. Only call G4 when those specific findings are documented.

User prompt prefix builder. The function `fs_prefix()` renders six worked examples, each in the block shown below, then appends the new-case instruction:

You will classify brain glioma cases from MRI reports. Below are worked examples from confirmed cases -- study the reasoning carefully:

```

=== EXAMPLE {i+1} ===
REPORT:
{rep[:1500]}

REASONING: {REASON_L1[l1]} {REASON_WG[wg]}

**level1={l1}**
**lgghgg={lh}**
**who_grade={wg}**

[... 6 examples ...]

===
Now classify the NEW case using the same reasoning-first format:

LABELS:
- level1: 0=GBM/IDH-WT, 1=IDH-mut diffuse, 2=Non-diffuse/pediatric, 3=Other/NEC
- lgghgg: 0=LGG, 1>HGG
- who_grade: 0=G1/G2, 1=G3, 2=G4

```

```

Output format (REASONING FIRST, then labels):
REASONING: <2-4 sentences identifying key imaging features and their classification
implications. Explicitly check for T2-FLAIR mismatch (-> IDH-mut), and for necrosis or
microvascular proliferation (-> G4 vs G3).>

**level1=N**
**lgghgg=N**
**who_grade=N**

REPORT:
{report_text}

```

Per-class reasoning templates substituted into each worked example by index:

```

REASON_L1 = {

```

```

0: 'GBM/IDH-WT (HGG): typically ring enhancement, central necrosis, mass effect, deep white
   matter location, age >55. T2-FLAIR mismatch is ABSENT.',
1: 'IDH-mut diffuse: T2-FLAIR mismatch sign present, less aggressive appearance, frontal/
   temporal cortex/subcortex, younger patient (<50). Often non-enhancing or minimal
   enhancement.',
2: 'Non-diffuse/pediatric: cystic+enhancing nodule (pilocytic) OR midline location (H3-
   altered). Lower aggression. Often pediatric/young adult.',
3: 'Other/NEC: atypical features -- extra-axial mass (meningioma-like), multiple foci (
   metastases), or features that do not fit a specific glioma class.',
}
REASON_WG = {
0: 'G1/G2 (low grade): minimal/no enhancement, less mass effect, no necrosis or
   microvascular proliferation.',
1: 'G3 (anaplastic): increased cellularity and atypia, may have mitoses, but NO necrosis
   and NO microvascular proliferation. Distinguishing G3 from G4 requires absence of these
   specific features.',
2: 'G4 (GBM-equivalent): necrosis OR microvascular proliferation present. Most aggressive
   radiographic appearance.',
}

```

The exemplar picker selects six exemplars from `train.csv` (two GBM, two IDH-mut diffuse, one Non-diffuse/pediatric, one Other/NEC), with a length-based score that prefers reports of 800–1500 characters; if the selection contains no WHO Grade 3 case it appends the highest-scoring G3 case from a separate G3 pool, as described in Section 3.

C. Reproducibility

The inference code and the trained model bundle are available on request and are hosted in the repository <https://github.com/AnubhavBharadwaj/mediqa-core-2026-brihadaranyaka> (access shared with the task organizers; available to reviewers on request). The repository provides a self-contained inference script (`inference.py`) that reproduces our predictions on unseen subjects under any modality-availability condition, the serialized model bundle (`model_bundle_path.a.pt`, released via Git LFS) containing the trained per-head weights and the Dirichlet combiner, and a `Dockerfile` pinning the runtime environment used for the organizers' code-verification re-execution. The script extracts each modality's features, applies the per-head classifiers and the Dirichlet combiner, and (when an API key is supplied) the confidence-gated LLM override, then writes integer per-prong labels. The CoRe-BT training, validation, and test data are governed by a University of Washington data use agreement and are not redistributed; they are obtained directly from the task organizers.