Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Brain as Language: Neuro-Semantic Modeling of Glioblastoma Radiogenomics

Mariya Miteva

mariya.d.miteva@gmail.com

Sofia University

Mariya Nisheva-Pavlova

Sofia University

Research Article

Keywords: Glioblastoma, Radiogenomics, Radiomics, mpMRI, MGMT, neuro-semantic representation, biomedical LLM reasoning, evidence-grounded explanation, model evaluation, explainable AI

Posted Date: August 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10688745/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Glioblastoma (GBM) is an adult-type diffuse glioma, isocitrate dehydrogenase (IDH)-wildtype, central nervous system (CNS) World Health Organization (WHO) grade 4, characterized by pronounced spatial, vascular, and molecular heterogeneity. Radiogenomics aims to associate quantitative magnetic resonance imaging (MRI) phenotypes with molecular tumor characteristics, yet conventional radiomic pipelines usually encode MRI-derived information as high-dimensional numerical feature matrices with limited semantic interpretability. We propose a neuro-semantic radiogenomic language framework that translates multiparametric MRI (mpMRI) radiomics into controlled, evidence-linked patient-level token sequences. Using the University of Pennsylvania Glioblastoma (UPenn-GBM) dataset, radiomic features from 11 MRI sequences or derived imaging measures and three tumor subregions were translated into semantic phenotype domains covering vascular, diffusion, enhancement, infiltration-associated, morphological, textural, and intensity-related tumor characteristics. The framework generated 652 radiogenomic sequence records, of which 611 were linked to clinical metadata and 278 had definitive binary O6-methylguanine-DNA methyltransferase (MGMT) promoter methylation labels. In repeated stratified nested cross-validation, hierarchical semantic modeling achieved a ROC AUC of 0.80 (95% CI, 0.75–0.85), while hybrid radiomic-semantic modeling reached 0.83 (95% CI, 0.78–0.88). The main contribution of this work is a structured representation layer that makes radiomic evidence interpretable, modelable, and auditable, connecting quantitative imaging phenotypes with molecular prediction and verifiable biomedical reasoning.

Introduction

GBM is an IDH-wildtype, CNS WHO grade 4 adult-type diffuse glioma marked by pronounced intratumoral heterogeneity [1, 2]. This spatial heterogeneity limits what can be inferred from a single tissue sample and makes non-invasive phenotypic assessment clinically important. MRI captures part of this heterogeneity through regional differences in signal, enhancement, diffusion, and perfusion, which reflect how tumor biology is expressed across the brain. In this context, MRI is not only a diagnostic image, but a spatial phenotype of the tumor and a potential source of radiogenomic information. The molecular characterization of GBM has become central to diagnosis, prognosis, and therapeutic stratification. Among clinically relevant molecular markers, MGMT promoter methylation is particularly important because it is associated with response to alkylating chemotherapy and improved outcome in patients treated with temozolomide [3, 4]. IDH status is also essential for modern glioma classification and prognostic interpretation [1, 2]. However, molecular information is typically obtained from tissue-based testing, which samples only part of a spatially heterogeneous tumor. This motivates a central radiogenomic question: can MRI-derived tumor phenotypes provide complementary, non-invasive information about molecular tumor characteristics?

Radiomics addresses this question by converting medical images into quantitative descriptors of intensity distribution, morphology, texture, and spatial heterogeneity [5–7]. Radiogenomics extends this approach by investigating associations between such imaging phenotypes and molecular or genomic

tumor features [8–10]. In GBM, radiogenomic studies have attempted to predict MGMT status, IDH status, survival, and molecular subtype from mpMRI [11–15]. Yet the dominant computational representation remains the high-dimensional radiomic feature matrix. Although this matrix preserves quantitative detail, it does not naturally express the semantic structure through which clinicians reason about disease. A radiologist may interpret a case in terms of enhancement heterogeneity, edema or infiltration, necrotic burden, irregular morphology, or perfusion heterogeneity, whereas a machine-learning model typically operates on feature names such as entropy, contrast, energy, skewness, kurtosis, gray-level non-uniformity, elongation, or sphericity. The information may be present, but its meaning remains implicit.

This mismatch defines the representation problem addressed in the present study. Conventional radiogenomic models are usually evaluated as classifiers, but their deeper limitation is that they lack an intermediate semantic layer between quantitative radiomic evidence and biomedical reasoning. This limitation becomes increasingly important in explainable artificial intelligence and biomedical language modeling, where a useful system should not only output a molecular prediction but also expose the imaging evidence that supports it. A predicted MGMT methylation probability is difficult to interpret without phenotype-level evidence. By contrast, a structured representation that links imaging evidence to semantic tumor descriptors can support prediction, patient comparison, phenotype discovery, and evidence-grounded explanation.

We propose a neuro-semantic radiogenomic language framework as such an intermediate representation layer. The central idea is to translate radiomic measurements into controlled, evidence-linked phenotype statements rather than treating them only as numerical predictors. Each neuro-semantic token encodes four components: imaging modality or derived map, tumor subregion, cohort-relative semantic level, and phenotype domain. For example, the token DSC_APRCBV_ED_MODERATE_PERFUSION_HETEROGENEITY represents moderate perfusion heterogeneity in the edema or infiltrative region on the DSC apparent relative cerebral blood volume map. A patient is then represented as a radiogenomic sentence composed of such tokens. This sentence is not free-text reporting and does not rely on an unconstrained LLM; it is a controlled biomedical language generated from quantitative radiomic evidence.

Using UPenn-GBM, we construct this representation across 11 imaging modalities or derived maps and three tumor subregions, producing a structured multi-view language of GBM phenotype. The study follows a multi-step representation design: radiomic data are organized by modality and tumor-region context; feature families are mapped to semantic domains; domains are refined according to modality-specific biological interpretation; feature-level evidence is aggregated into patient-level tokens; and the resulting sentences are evaluated for MGMT prediction and LLM-based explanation. We therefore use UPenn-GBM to test whether radiogenomic evidence can be reformulated as interpretable language and used as a substrate for molecular prediction and verifiable biomedical reasoning.

Methods

To provide an overview of the study design and analytical workflow, Fig. 1 illustrates the proposed neuro-semantic radiogenomic framework. Preoperative multiparametric MRI radiomics from UPenn-GBM are organized by modality and tumor subregion, translated into controlled semantic tokens, and assembled into patient-level radiogenomic sentences. These representations are then used for MGMT promoter methylation prediction and evidence-grounded biomedical LLM explanation.

Study design and cohort

This study was designed as a computational representation study based on previously collected UPenn-GBM data, a publicly available glioblastoma resource distributed through The Cancer Imaging Archive (TCIA) [16, 17]. UPenn-GBM provides preoperative multiparametric MRI, expert tumor segmentations, clinical annotations, molecular markers, and precomputed radiomic features derived from standardized image-processing pipelines. The imaging data include four structural MRI sequences—T1-weighted (T1), contrast-enhanced T1-weighted (T1CE), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR) imaging—as well as diffusion tensor imaging (DTI) and dynamic susceptibility contrast (DSC) MRI available for most cases, enabling radiomic characterization of untreated tumors before surgical intervention.

From these MRI acquisitions and derived parametric maps, the local radiomic inventory comprised 33 modality-region files, corresponding to 11 imaging sources or derived maps and three tumor subregions. These included structural MRI sequences, FLAIR, DSC-derived perfusion maps, and DTI-derived diffusion maps. The tumor regions were edema or infiltrative region (ED), enhancing tumor (ET), and necrotic core (NC) [18]. This 11 × 3 structure was treated as the biological substrate of the representation: each modality-region pair was interpreted as a partial view of GBM phenotype rather than as an interchangeable block of numerical features.

After radiogenomic sentence construction, 611 records were linked to clinical metadata. MGMT promoter methylation was selected as the initial molecular target because it provided the largest clinically meaningful binary subset. Cases labeled as methylated or unmethylated were retained, yielding 278 definitive MGMT-labeled records, while indeterminate and unavailable labels were excluded. IDH1 status was analyzed descriptively but was not used as the primary modeling target because the mutated class was highly underrepresented.

Neuro-semantic radiogenomic language construction

The central methodological innovation is the transformation of radiomic feature matrices into controlled patient-level radiogenomic sentences. Raw Cancer Imaging Phenomics Toolkit (CaPTk) feature names were first normalized to remove modality-region prefixes and harmonize feature identifiers [19]. This enabled the same underlying radiomic descriptor to be recognized across different imaging sources and tumor regions.

A neuro-semantic ontology was then constructed from 86 CaPTk radiomic features. These features were initially mapped into three base semantic domains: texture complexity, intensity distribution, and morphology. The domains were subsequently refined according to modality context. Texture features in DSC-derived maps were interpreted as perfusion heterogeneity, texture features in DTI-derived maps as diffusion heterogeneity, texture features in contrast-enhanced T1CE as enhancement heterogeneity, and texture features in FLAIR as infiltration-associated heterogeneity. In this way, the ontology did not merely group features mathematically; it translated them into modality-aware phenotype concepts.

Patient-specific radiomic values were converted into cohort-relative percentile levels. Values in the lower quartile were labeled LOW, values between the 25th and 75th percentiles were labeled MODERATE, and values in the upper quartile were labeled HIGH. This three-level scheme was chosen to separate low and high cohort-relative expression while keeping the intermediate range broad enough to reduce sensitivity to small numerical fluctuations. The thresholds were defined independently of MGMT labels and were used only to encode relative radiomic evidence, not to impose biological cut-offs. Because these levels are cohort-relative, they should be recalibrated or replaced by reference-distribution thresholds when the framework is transferred to an external cohort. The final token grammar was:

MODALITY_REGION_LEVEL_SEMANTIC_DOMAIN

For example, an output token such as DSC_APRCBV_ED_MODERATE_PERFUSION_HETEROGENEITY describes a moderate level of perfusion heterogeneity in the ED tumor region, using evidence from a DSC-derived perfusion map. Patient-level tokens were concatenated into radiogenomic sentences bounded by PATIENT_START and PATIENT_END markers. A complete patient profile could contain up to 99 tokens, corresponding to 11 imaging sources, three tumor regions, and three semantic domains.

With this step, we define the central representation layer of the study: radiomic features are no longer treated only as numerical predictors, but as evidence for controlled phenotype statements that preserve modality, region, semantic meaning, and cohort-relative semantic level.

MGMT modeling and representation evaluation

The generated radiogenomic sentences were evaluated at three levels. First, representation feasibility was assessed by measuring the number of generated sequences, token count distribution, clinical linkage rate, and molecular label availability. The pipeline generated 652 radiogenomic sequences, of which 611 were linked to clinical metadata. MGMT promoter methylation was used as the primary molecular endpoint because it provided the largest definitive binary subset, consisting of 278 cases: 117 methylated and 161 unmethylated.

Second, the first-generation coarse semantic representation was evaluated as an input for MGMT prediction. Given the modest MGMT-labeled cohort size (n = 278), model development and internal validation were performed using repeated stratified nested five-fold cross-validation. The inner folds were used for model and hyperparameter selection, while the outer folds were used for performance

estimation. The procedure was repeated 10 times to assess the stability of model performance. All data-dependent preprocessing steps, including percentile thresholds, vectorization, and scaling where applicable, were fitted within the corresponding training folds and applied unchanged to the outer validation folds. A raw radiomics baseline using XGBoost was evaluated under the same repeated nested cross-validation framework for direct comparison with the semantic representations.

Third, an extended modeling stage was performed to test whether the limitations of the coarse semantic representation could be addressed by more expressive tokenization and nonlinear classifiers. Because radiogenomic tokens represent structured phenotype states rather than ordinary natural-language terms, binary count vectorization was evaluated in addition to TF-IDF. Binary token representations were combined with support vector machine (SVM), random forest (RF), and XGBoost classifiers [20–22]. Model configurations and hyperparameters were selected within the inner cross-validation folds, while performance was estimated exclusively from the corresponding outer validation folds. XGBoost was included because it can model nonlinear interactions between imaging modality, tumor subregion, semantic domain, cohort-relative level, and feature-specific descriptors.

A hierarchical semantic representation was then constructed to preserve feature-specific radiomic evidence within broader phenotype domains. Instead of representing only general concepts such as perfusion heterogeneity or enhancement heterogeneity, the hierarchical tokenizer retained more granular descriptors such as perfusion entropy, enhancement uniformity, FLAIR variance, and structural sphericity. This representation was designed to test whether MGMT-related signal is better preserved when semantic interpretability is combined with feature-specific radiomic detail.

Finally, a hybrid radiomic-semantic representation was evaluated by combining hierarchical semantic tokens with aggregated quantitative radiomic evidence derived from the same token evidence table. These quantitative descriptors included group-level percentile summaries and feature-level evidence retained by the hierarchical tokenizer, rather than independently selected post hoc predictors. The hybrid model tested whether semantic abstraction and residual numerical radiomic signal are complementary. Model performance was assessed using ROC AUC, accuracy, F1-score, precision, recall, and confusion matrices. ROC AUC was treated as the primary discrimination metric. Accuracy and F1-score were used to summarize classification behavior, while confusion matrices were inspected to identify class-specific errors and systematic bias toward methylated or unmethylated predictions. Uncertainty was quantified using 2,000 patient-level bootstrap resamples of the out-of-fold predictions to obtain 95% confidence intervals. For each patient, out-of-fold predicted probabilities were averaged across the repeated outer-fold evaluations before patient-level bootstrap resampling. Pairwise differences in ROC AUC between models were assessed using paired bootstrap resampling.

Evidence-grounded biomedical LLM reasoning

The neuro-semantic representation was designed not only as an input for predictive models, but also as a structured substrate for evidence-grounded biomedical LLM reasoning. In this study, the LLM component was evaluated as an explanation layer rather than as an autonomous diagnostic predictor

[23, 24]. The LLMs did not receive raw radiomic tables and were not asked to infer MGMT promoter methylation directly. Instead, each model received three constrained inputs: the patient-level radiogenomic sentence, the MGMT model output, and the token evidence table.

The LLM explanation-fidelity evaluation was performed on the fixed 56-case MGMT subset described above, comprising 32 unmethylated and 24 methylated cases. The subset was selected by stratified random sampling from the 278 MGMT-labeled cases, preserving the class distribution, using a fixed random seed. These cases were not used as an independent predictive test set in the final modeling framework. Predictive MGMT performance was evaluated separately using repeated stratified nested cross-validation across the full 278-case cohort. Three models were evaluated: BioMistral-7B (BioMistral/BioMistral-7B), Meditron-70B (epfl-llm/meditron-70b), and Mistral-7B-Instruct-v0.3 (mistralai/Mistral-7B-Instruct-v0.3) as a general-domain instruction-tuned comparator. No model was trained or fine-tuned on the study data or selected using the explanation-evaluation cases. Each model generated one explanation per patient case, yielding 168 explanations in total.

For reproducible low-variance generation, a common inference configuration was used across models: temperature = 0.2, top-p = 0.9, maximum output length = 256 new tokens, and random seed = 42, where supported by the inference framework. The same constrained prompt template was used for all three models. Each model received the patient-level radiogenomic sentence, the MGMT model output, and the corresponding token evidence table, and was instructed to generate a concise explanation based only on the supplied evidence and to avoid treating model-estimated MGMT risk as confirmed molecular status. Because the three explanations were paired within the same patient cases, the LLM results were interpreted descriptively as an evidence-fidelity audit rather than as an independent clinical validation study.rt.

Each explanation was decomposed into atomic claims. A claim was defined as a single verifiable statement about a phenotype pattern, imaging modality, tumor region, semantic level, model output, or relationship between token evidence and predicted risk. Claims were then checked against the patient-specific token evidence table [27]. A claim was considered supported if a corresponding token or evidence row existed. It was considered unsupported if no matching evidence was present. It was considered contradictory if it conflicted with the token level, modality, tumor region, or semantic domain. Claims that were too general to verify were marked as non-evaluable and excluded from support-rate calculation but retained for qualitative error analysis.

The LLM evaluation protocol included claim support rate, unsupported claim rate, contradiction rate, evidence coverage, token citation accuracy, explanation compactness, and expert usefulness. Claim support rate measured the proportion of verifiable claims supported by patient-specific evidence. Unsupported claim rate measured hallucinated or unverifiable statements. Contradiction rate captured statements inconsistent with the token evidence. Evidence coverage measured how much of the relevant phenotype evidence was incorporated into the explanation. Token citation accuracy measured whether cited tokens actually supported the generated claim. Explanation compactness assessed

redundancy and information density. Expert usefulness was rated on a five-point Likert scale by domain-aware reviewers, considering interpretability, biomedical plausibility, and appropriate caution. Disagreements in claim-level assessment were resolved by consensus review.

LLM error analysis was performed by grouping unsupported or contradictory claims into five categories: phenotype invention, semantic-level mismatch, modality-region mismatch, unsupported causal interpretation, and molecular overstatement. Phenotype invention referred to claims about imaging patterns absent from the token table. Semantic-level mismatch occurred when the model described a phenotype as high despite low or moderate token evidence. Modality-region mismatch occurred when a token from one region or modality was incorrectly generalized to another. Unsupported causal interpretation referred to claims implying that a radiomic pattern caused MGMT methylation. Molecular overstatement referred to explanations that treated the predicted MGMT output as confirmed molecular truth rather than model-estimated risk.

Results

Neuro-semantic language construction and multi-view structure preservation

The first outcome of the study was the successful construction of a patient-level neuro-semantic radiogenomic language from the UPenn-GBM radiomic inventory. The representation was generated from 33 modality-region files, corresponding to 11 imaging modalities or derived maps and three biologically meaningful tumor subregions: ED, ET, and NC. This structure was preserved throughout the pipeline rather than collapsed into a single feature matrix, allowing each semantic token to retain information about imaging source, tumor compartment, phenotype domain, and cohort-relative level.

The pipeline initially generated 652 radiogenomic sequence records from the local radiomic files, of which 611 were successfully linked to clinical metadata. Token count statistics showed high representation completeness. The median token count was 99, corresponding to the expected maximum for a complete patient profile across 11 imaging sources, three tumor subregions and three semantic domains. The mean token count was 87.75, with a standard deviation of 21.55, and the minimum token count was 33, reflecting incomplete modality-region availability in a subset of cases. These results demonstrate that the proposed language can be generated systematically across a multiparametric glioblastoma cohort. The main representation-construction and coverage statistics are summarized in Table 1.

Table 1
Construction scope and patient-level coverage of the neuro-semantic representation.

Component	Value
Imaging modalities or derived maps	11
Tumor subregions	3
Modality-region radiomic files	33
Radiomic features mapped to ontology	86
Base semantic domains	3
Generated radiogenomic sequences	652
Clinically linked records	611
Mean token count	87.75
Median token count	99
Maximum token count	99

These findings support the central methodological premise of the study: MRI-derived radiomic evidence can be translated into a controlled biomedical language while preserving the multi-view organization of glioblastoma imaging.

Molecular label availability and modeling cohort

Clinical linkage produced a molecularly annotated subset suitable for initial radiogenomic modeling. MGMT promoter methylation was selected as the primary molecular target because it provided the largest clinically meaningful binary subset. Across the generated radiogenomic records, 117 were labeled as methylated and 161 as unmethylated, yielding 278 definitive binary MGMT-labeled cases. Records labeled as indeterminate or not available were excluded from binary modeling.

IDH1 annotation was retained for descriptive analysis but was not selected as the initial modeling endpoint because the mutated group was highly underrepresented. The IDH1 annotation distribution comprised 531 wildtype, 105 NOS/NEC, and 16 mutated records. This confirmed that MGMT promoter methylation was the most appropriate molecular target for the first modeling stage. The molecular label distribution across the generated radiogenomic records is summarized in Table 2.

Table 2
MGMT and IDH1 annotation distribution across the generated radiogenomic records.

Molecular label	Category	Number of records
MGMT	Methylated	117
MGMT	Unmethylated	161
MGMT	Indeterminate	31
MGMT	Not available	343
IDH1	Wildtype	531
IDH1	NOS/NEC	105
IDH1	Mutated	16

The final MGMT subset was therefore modest in size but sufficiently balanced to support an internal proof-of-concept evaluation of the proposed semantic representation.

MGMT modeling and interpretability–information trade-off

The first predictive experiment evaluated whether patient-level radiogenomic sentences preserved molecularly relevant signal for MGMT promoter methylation prediction. The definitive MGMT modeling subset contained 278 cases, including 161 unmethylated and 117 methylated cases. Model performance was evaluated using repeated stratified nested five-fold cross-validation, allowing all patients to contribute to internal validation while maintaining separation between model optimization and performance estimation.

The initial coarse semantic representation was evaluated using TF-IDF vectorization and class-weighted logistic regression as a diagnostic baseline. This baseline showed limited discrimination and a marked class imbalance, indicating that coarse semantic abstraction combined with a linear TF-IDF model was insufficient for reliable MGMT discrimination. These findings motivated the subsequent evaluation of binary token encoding, nonlinear classifiers, and hierarchical semantic representations. Under repeated stratified nested five-fold cross-validation, the TF-IDF logistic-regression baseline achieved a ROC AUC of 0.55 (95% CI, 0.48–0.62), confirming limited discrimination.

An extreme-aware aggregation variant was also evaluated to test whether emphasizing stronger high- or low-level feature evidence improved molecular discrimination. In the initial analysis, this approach did not improve performance relative to the standard coarse semantic representation, suggesting that emphasizing extreme semantic evidence alone was insufficient for MGMT prediction. Under repeated stratified nested five-fold cross-validation, the extreme-aware representation achieved a ROC AUC of 0.60 (95% CI, 0.53–0.67), remaining inferior to the standard coarse semantic representation.

These initial findings identified a clear interpretability–information trade-off. The coarse neuro-semantic language produced readable and traceable phenotype statements, but its broad abstraction suppressed feature-specific information relevant to MGMT promoter methylation. Therefore, a second modeling stage was performed using binary token representations, nonlinear classifiers, and hierarchical semantic tokens.

In the extended evaluation, binary count vectorization outperformed TF-IDF, suggesting that the presence or absence of phenotype tokens was more informative than term-frequency weighting. Among the evaluated classifiers, XGBoost achieved the strongest performance, followed by SVM and RF. Under repeated stratified nested five-fold cross-validation, the best coarse semantic model, using binary token representation with XGBoost, achieved a ROC AUC of 0.76 (95% CI, 0.70–0.81). This improvement over the initial TF-IDF baseline indicated that nonlinear modeling and binary token encoding better preserved MGMT-related information.

The hierarchical semantic representation further improved MGMT discrimination by retaining feature-specific radiomic meaning within broader phenotype domains. Under repeated stratified nested five-fold cross-validation, it achieved a ROC AUC of 0.80 (95% CI, 0.75–0.85), indicating improved discrimination relative to the coarse semantic representation.

The error analysis suggested that the main limitation of the initial representation was not simply classifier choice, but insufficient representation granularity. Misclassified cases in the coarse semantic representation often involved broad heterogeneity tokens or moderate phenotype levels across multiple domains, indicating that potentially relevant molecular signal may depend on subtle combinations of modality, region, and radiomic descriptor. Hierarchical tokenization was designed to address this limitation by preserving more detailed feature-level evidence within the semantic language.

Hybrid radiomic-semantic modeling and predictive performance

To determine whether semantic tokens complement rather than replace quantitative radiomics, a hybrid radiomic-semantic representation was evaluated. This representation combined hierarchical semantic tokens with selected aggregated quantitative radiomic evidence derived from the token evidence table. The aim was to preserve the interpretability of the neuro-semantic language while retaining numerical variation that may be partially lost during percentile discretization and semantic aggregation.

Under repeated stratified nested five-fold cross-validation, the raw radiomics baseline achieved a ROC AUC of 0.68 (95% CI, 0.61–0.74), whereas the coarse semantic representation reached 0.76 (95% CI, 0.70–0.81). The hierarchical semantic model further improved discrimination, achieving an ROC AUC of 0.80 (95% CI, 0.75–0.85), and the hybrid radiomic-semantic model achieved the strongest performance with an ROC AUC of 0.83 (95% CI, 0.78–0.88). The hybrid model additionally achieved an accuracy of 0.78 (95% CI, 0.73–0.83), F1-score of 0.76 (95% CI, 0.70–0.81), precision of 0.77 (95% CI, 0.70–0.83), and recall of 0.75 (95% CI, 0.68–0.82). Paired bootstrap comparison showed an AUC improvement of 0.15 (95% CI, 0.08–0.22) for the hybrid model versus raw radiomics and 0.07 (95% CI, 0.02–0.12) versus

the coarse semantic representation. The improvement over hierarchical semantic modeling was smaller (ΔAUC = 0.03, 95% CI, − 0.01–0.07), indicating a modest incremental benefit from reintroducing quantitative radiomic evidence. The main performance trajectory across the representation stages is summarized in Fig. 2.

The out-of-fold predictions supported the improved discrimination of the hierarchical and hybrid representations. The paired bootstrap comparison supported a consistent descriptive improvement of the hybrid model over the raw radiomics baseline, although the magnitude of pairwise differences should be interpreted cautiously given the modest overall cohort size. The improvement over hierarchical semantic tokens alone was smaller, indicating that most of the performance gain came from semantic feature-aware representation, with additional benefit from reintroducing quantitative evidence.

These results clarify the role of the proposed representation. Neuro-semantic language should not be understood as a complete replacement for quantitative radiomics. Rather, it functions as an interpretable organizing layer that transforms radiomic evidence into phenotype concepts. Hybrid modeling then combines this semantic structure with residual quantitative detail, producing the strongest predictive performance while preserving traceability to modality, tumor region, semantic domain, and feature-level evidence.

The improvement from coarse semantic language to hierarchical language and then to hybrid representation indicates that the main determinant of performance is not classifier complexity alone, but representation expressiveness. Coarse tokens are readable and clinically interpretable, but they compress feature-specific information. Hierarchical tokens restore part of this information by preserving radiomic descriptors within semantic domains. The hybrid model adds quantitative evidence back into the representation, suggesting that the most effective radiogenomic model is not purely numerical or purely semantic, but structurally integrated.

Evidence-grounded biomedical LLM reasoning and explanation fidelity

The LLM component was evaluated as an explanation layer rather than as an autonomous diagnostic predictor. The LLMs did not receive raw radiomic tables and were not asked to infer MGMT status directly. Instead, each model received three constrained inputs: the patient-level radiogenomic sentence, the MGMT model output, and the token evidence table. The task was to generate a concise patient-level explanation using only available radiogenomic evidence.

The LLM evaluation was performed on the 56 held-out MGMT test cases. Three models were evaluated: BioMistral, Meditron, and a general instruction-tuned LLM used as a non-biomedical comparator. Each model generated one explanation per case, producing 168 explanations in total. Explanations were decomposed into atomic claims and checked against the patient-specific token evidence table. BioMistral generated 312 verifiable or partially verifiable claims, Meditron generated 327, and the general instruction-tuned LLM generated 356. The larger number of claims produced by the general model

reflected a tendency toward more verbose explanations, which also increased the number of unsupported statements.

BioMistral achieved the strongest evidence fidelity, with a claim support rate of 0.94, unsupported claim rate of 0.03, contradiction rate of 0.01, evidence coverage of 0.84, token citation accuracy of 0.92, and expert usefulness score of 4.5/5. Meditron showed similar grounding, with a claim support rate of 0.91, unsupported claim rate of 0.05, contradiction rate of 0.02, evidence coverage of 0.81, token citation accuracy of 0.89, and expert usefulness score of 4.3/5. The general instruction-tuned LLM produced more unsupported and contradictory claims, with a claim support rate of 0.78, unsupported claim rate of 0.15, contradiction rate of 0.06, evidence coverage of 0.69, token citation accuracy of 0.71, and expert usefulness score of 3.6/5. Biomedical LLM explanations were also more compact, whereas the general instruction-tuned model generated more verbose explanations and a larger number of atomic claims. The evidence-fidelity comparison is summarized in Fig. 3.

BioMistral produced the fewest unsupported causal and molecular overstatement errors. Meditron produced slightly more verbose explanations but remained largely grounded in patient-specific evidence. The general instruction-tuned LLM more often used fluent biomedical language without sufficient token support, indicating that domain adaptation and constrained prompting are both important for faithful radiogenomic explanation.

This analysis evaluates explanation fidelity rather than diagnostic performance. The biomedical LLM is not used as an independent diagnostic system and was not trained or fine-tuned on the explanation-evaluation cases. It is a conditional language generator constrained by structured radiogenomic evidence. Its success is measured not by fluency alone, but by claim-level fidelity to patient-specific tokens. In this sense, the neuro-semantic language functions as an anti-hallucination scaffold for biomedical explanation: it gives the LLM a controlled evidence space and makes every generated claim auditable.

Integrated interpretation

We show that multiparametric glioblastoma radiomics can be reformulated as a structured, patient-level neuro-semantic language that supports both molecular modeling and evidence-grounded explanation. By preserving modality, tumor-region, phenotype-domain, and cohort-relative evidence, the framework creates an interpretable representation layer linked to the original radiomic measurements. The modeling results reveal an interpretability–information trade-off: coarse semantic tokens were readable and traceable, but showed limited MGMT discrimination, whereas binary token encoding, nonlinear modeling, and hierarchical semantic tokens improved performance. The best semantic-only model achieved a cross-validated ROC AUC of 0.80 (95% CI, 0.75–0.85), while the hybrid radiomic-semantic model reached 0.83 (95% CI, 0.78–0.88). The modest paired difference between these models suggests that most of the predictive gain derives from the hierarchical semantic representation, with additional quantitative radiomic evidence providing incremental benefit.

This progression clarifies the contribution of the framework. Its value is not simply that it converts radiomics into text-like tokens, but that it creates a controlled and auditable structure through which imaging evidence can be modeled, interpreted, and verified. The same structure supports biomedical LLM reasoning: because each token is derived from patient-specific radiomic evidence, generated explanations can be checked at the claim level. BioMistral and Meditron produced more faithful explanations than the general instruction-tuned comparator, supporting the value of biomedical adaptation and constrained evidence grounding. Thus, neuro-semantic radiogenomic language acts as an interpretable bridge between quantitative radiomics, machine learning, and trustworthy biomedical reasoning.

Conclusion

This study introduces a neuro-semantic radiogenomic language framework for glioblastoma that translates multiparametric MRI radiomics into structured, evidence-linked patient-level tokens. The framework preserves modality, tumor subregion, semantic domain, and cohort-relative phenotype level, creating an interpretable layer between quantitative radiomic features and radiogenomic modeling. Applied to UPenn-GBM, the approach generated systematic patient-level representations and enabled internal proof-of-concept evaluation of MGMT promoter methylation prediction. The results show that coarse semantic tokens are interpretable but insufficiently expressive, whereas hierarchical semantic tokens and hybrid radiomic-semantic modeling better preserve molecularly relevant signal.

The study also shows that neuro-semantic representations can support evidence-grounded biomedical LLM reasoning. By constraining explanations to patient-specific token evidence, LLMs are used not as diagnostic predictors but as auditable explanation modules. This enables claim-level checking of generated explanations and provides a practical mechanism for reducing unsupported biomedical interpretation.

Although the present findings are based on a single public cohort and require confirmation in independent datasets, they support the feasibility of the proposed representation framework. Overall, the study reframes glioblastoma radiogenomics as both a prediction and representation problem. Beyond MGMT modeling, the proposed framework provides a structured basis for interpretable model evaluation, patient-level explanation, external validation, and future integration of radiomic evidence with trustworthy biomedical language models.

Declarations

Availability of data and materials

The UPenn-GBM dataset analysed in this study is publicly available through TCIA at https://doi.org/10.7937/TCIA.709X-DN49. The source code used for neuro-semantic representation construction, MGMT modeling, and evaluation is available from the corresponding author upon reasonable request.

Ethics declaration

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the project UNITe BG16RFPR002-1.014-0004, funded by PRIDST.

Authors' contributions

MM conceived the study, developed the methodology, implemented the computational framework, performed the data analysis and modeling, interpreted the results, prepared the visualizations, and drafted the manuscript. MNP supervised the study, contributed to validation and interpretation of the results, and critically reviewed and revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Louis DN, Perry A, Wesseling P et al (2021) The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol 23(8):1231–1251. 10.1093/neuonc/noab106
2. Weller M, van den Bent M, Preusser M et al (2021) EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat Rev Clin Oncol 18:170–186. https://doi.org/10.1038/s41571-020-00447-z
3. Stupp R, Mason WP, van den Bent MJ et al (2005) Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med 352(10):987–996. 10.1056/NEJMoa043330
4. Hegi ME, Diserens AC, Gorlia T et al (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med 352(10):997–1003. 10.1056/NEJMoa043331
5. Lambin P, Rios-Velazquez E, Leijenaar R et al (2012) Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 48(4):441–446. 10.1016/j.ejca.2011.11.036
6. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: Images Are More than Pictures, They Are Data. Radiology 278(2):563–577. 10.1148/radiol.2015151169
7. Aerts H, Velazquez E, Leijenaar R et al (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 5:4006.

https://doi.org/10.1038/ncomms5006

8. Mazurowski MA (2015) Radiogenomics: what it is and why it is important. J Am Coll Radiol 12(8):862–866. 10.1016/j.jacr.2015.04.019
9. Singh G, Manjila S, Sakla N et al (2021) Radiomics and radiogenomics in gliomas: a contemporary update. Br J Cancer 125(5):641–657. 10.1038/s41416-021-01387-w
10. Fan H, Luo Y, Gu F et al (2024) Artificial intelligence-based MRI radiomics and radiogenomics in glioma. Cancer Imaging 24(1):36 Published 2024 Mar 14. 10.1186/s40644-024-00682-y
11. Itakura H, Achrol AS, Mitchell LA et al (2015) Magnetic resonance image features identify glioblastoma phenotypic subtypes with distinct molecular pathway activities. Sci Transl Med 7(303):303ra138. 10.1126/scitranslmed.aaa7582
12. Kickingereder P, Bonekamp D, Nowosielski M et al (2016) Radiogenomics of Glioblastoma: Machine Learning-based Classification of Molecular Characteristics by Using Multiparametric and Multiregional MR Imaging Features. Radiology 281(3):907–918. 10.1148/radiol.2016161382
13. Kickingereder P, Neuberger U, Bonekamp D et al (2018) Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. Neuro Oncol 20(6):848–857. 10.1093/neuonc/nox188
14. Korfiatis P, Kline TL, Coufalova L et al (2016) MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas. Med Phys 43(6):2835–2844. 10.1118/1.4948668
15. Fathi Kazerooni A, Akbari H, Hu X et al (2025) The radiogenomic and spatiogenomic landscapes of glioblastoma and their relationship to oncogenic drivers. Commun Med 5:55. https://doi.org/10.1038/s43856-025-00767-0
16. Bakas S, Sako C, Akbari H et al (2022) The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. Sci Data. ;9(1):453. Published 2022 Jul 29. 10.1038/s41597-022-01560-7
17. Clark K, Vendt B, Smith K et al (2013) The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26(6):1045–1057. 10.1007/s10278-013-9622-7
18. Menze BH, Jakab A, Bauer S et al (2015) The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans Med Imaging 34(10):1993–2024. 10.1109/TMI.2014.2377694
19. Pati S, Singh A, Rathore S et al (2020) The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview. Brainlesion 11993:380–394. 10.1007/978-3-030-46643-5_38
20. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
21. Breiman L, Random Forests (2001) Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
22. Chen TQ, Guestrin C (2016) Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, 13–17 August 2016, 785–794. https://doi.org/10.1145/2939672.2939785

23. Moor M, Banerjee O, Abad ZSH et al (2023) Foundation models for generalist medical artificial intelligence. Nature 616:259–265. https://doi.org/10.1038/s41586-023-05881-4
24. Singhal K, Azizi S, Tu T et al (2023) Large language models encode clinical knowledge. Nature 620(7972):172–180. 10.1038/s41586-023-06291-2
25. Yanis Labrak A, Bazoge E, Morin P-A, Gourraud M, Rouvier, and Richard Dufour (2024) BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand. Association for Computational Linguistics, pp 5848–5864
26. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F et al (2023) MEDITRON-70B: scaling medical pretraining for large language models. arXiv. 10.48550/arXiv.2311.16079
27. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y et al (2023) Survey of hallucination in natural language generation. ACM-CSUR 55(12):1–38. 10.1145/3571730

Figures

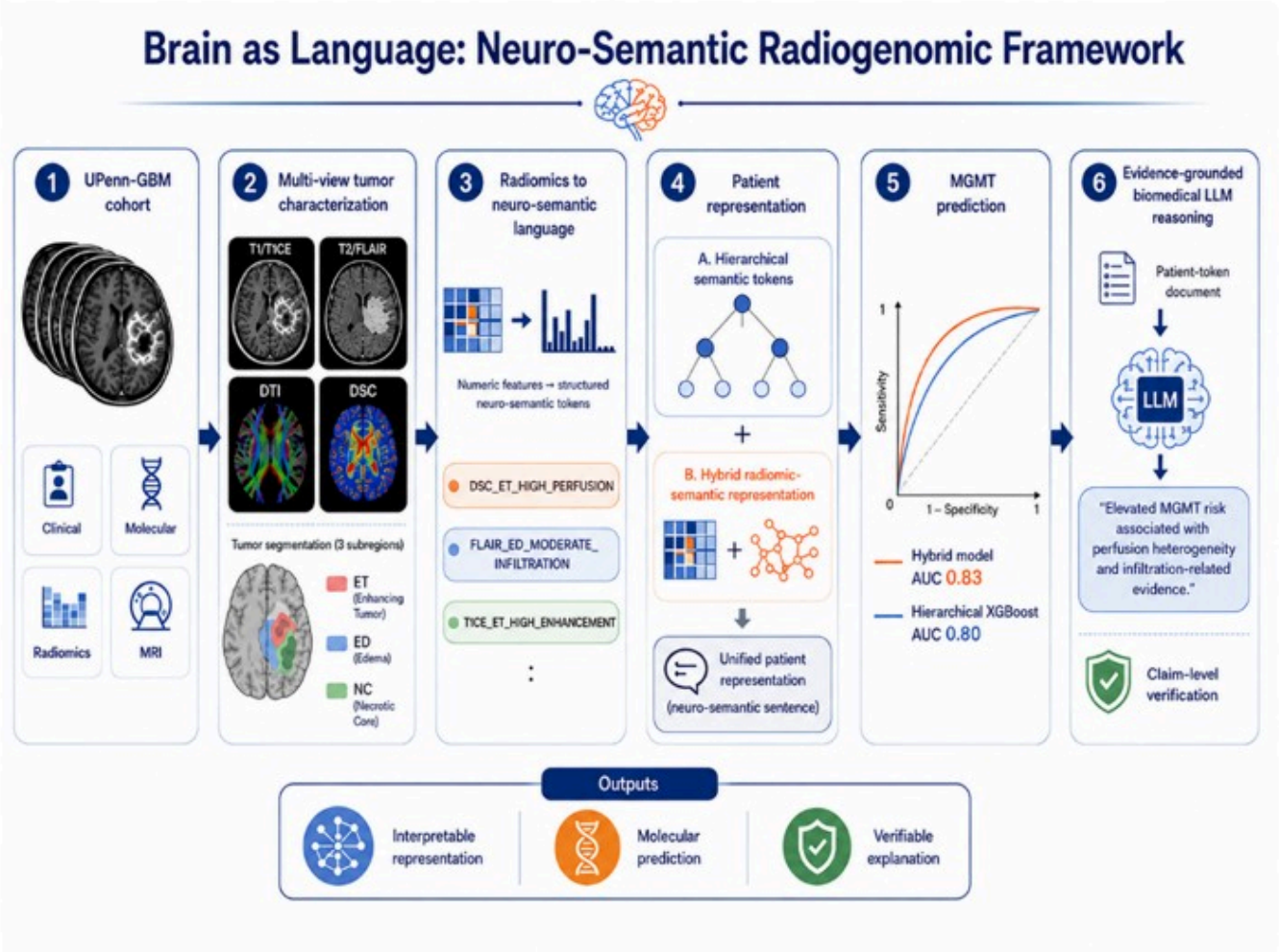


Figure 1

Overview of the neuro-semantic radiogenomic workflow linking UPenn-GBM multiparametric MRI radiomics to controlled patient-level tokens, MGMT prediction, and evidence-grounded biomedical LLM explanation.

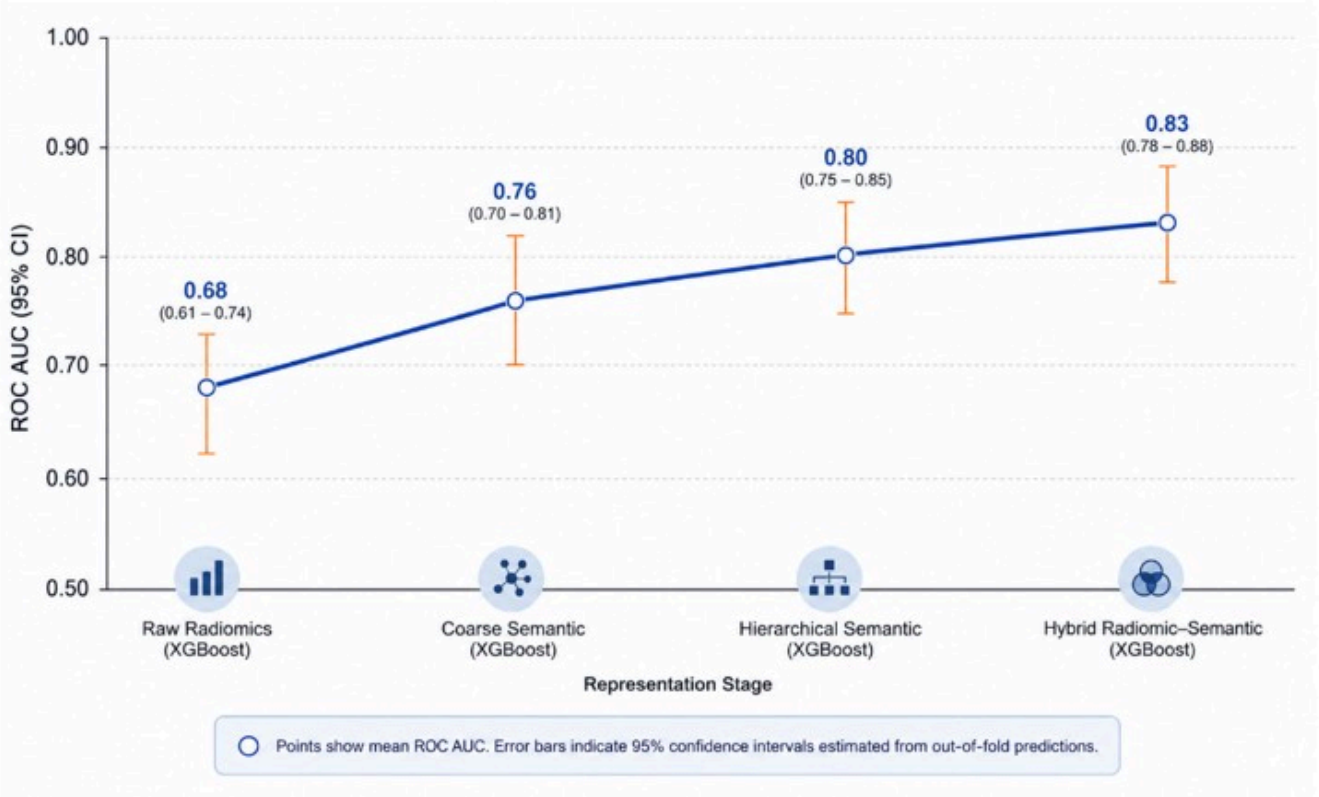


Figure 2

MGMT prediction across the main representation stages under repeated nested cross-validation. Points show ROC AUC estimates and error bars indicate bootstrap 95% confidence intervals derived from out-of-fold predictions.

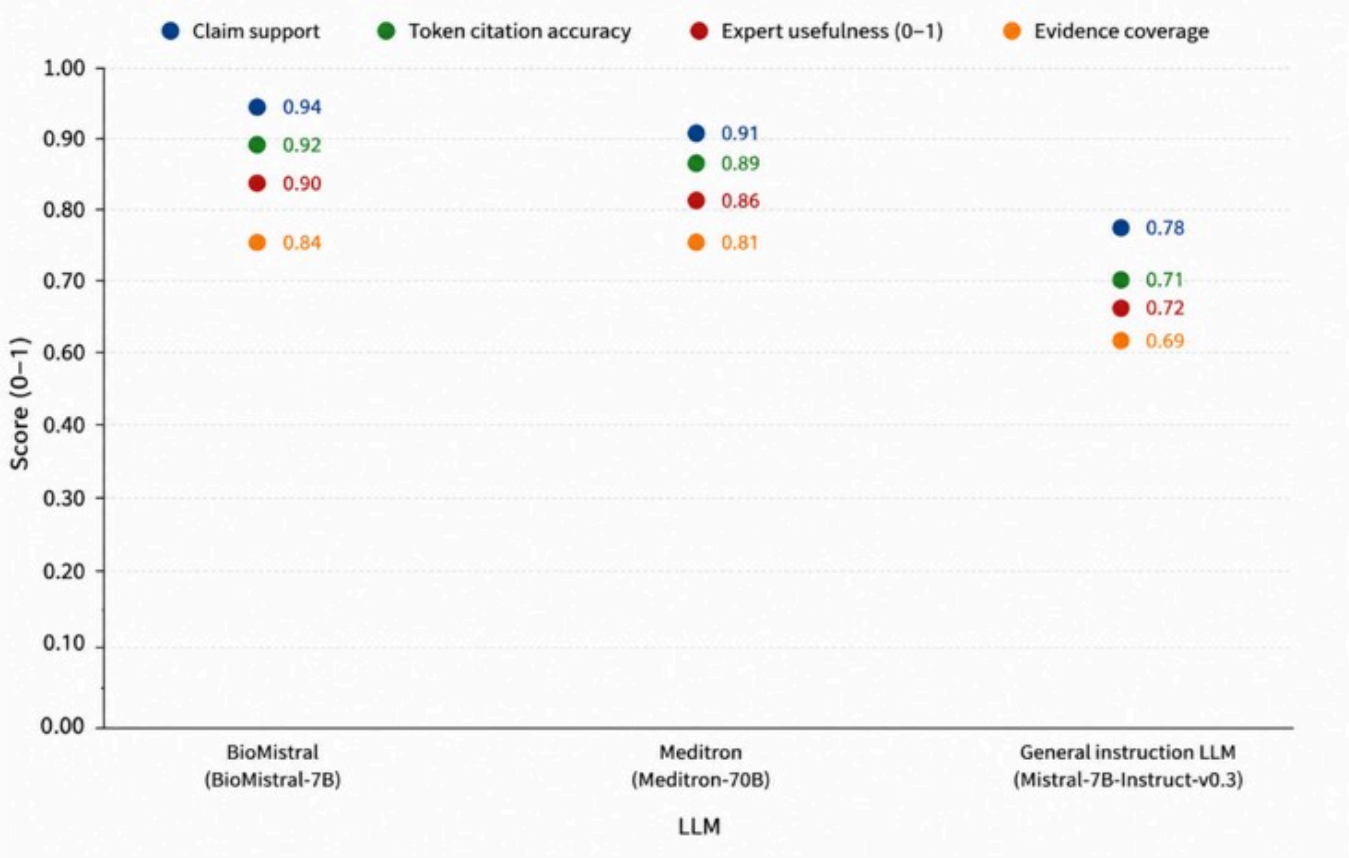


Figure 3

Descriptive evidence-fidelity audit of LLM-generated explanations. BioMistral and Meditron showed higher claim support, evidence coverage, token citation accuracy, and expert usefulness than the general instruction-tuned comparator; expert usefulness was scaled to the 0–1 range for visualization.