Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.264016

ARTIFICIAL INTELLIGENCE AND INFORMATICS

ORIGINAL ARTICLE

Glass-box agentic-style workflow for multiclass cine cardiac magnetic resonance imaging classification with a large language model

Ismail Mese[1]
Burak Kocak[2]

[1]Üsküdar State Hospital, Department of Radiology, İstanbul, Türkiye

[2]Basaksehir Cam and Sakura City Hospital, Department of Radiology, İstanbul, Türkiye

PURPOSE

To develop and evaluate a glass-box, agentic-style radiology pipeline that separates perception from reasoning for auditable multiclass diagnosis on cine cardiac magnetic resonance imaging (MRI), and to quantify accuracy, robustness across decoding temperatures, and fidelity/safety of generated narrative explanations.

METHODS

Using the labeled Automated Cardiac Diagnosis Challenge training cohort (n = 100; five diagnostic classes), cine bSSFP images were segmented at end-diastole (ED) and end-systole (ES) with a pretrained nnU-Net, and 17 clinically interpretable biomarkers were extracted. A large language model (LLM) (GPT-OSS-120B) queried prompts under three different prompt strategies (V1–V3) with majority-vote self-consistency after a stratified split into prompt development (n = 20) and independent evaluation (n = 80). Temperatures (T = 0.1, 1.0, and 2.0) were tested for stability. A decoupled narrative module generated radiologist-style reports. Narratives underwent radiologist audit for numeric fidelity and clinical safety. Machine learning algorithms [Random Forest, Support Vector Machine (SVM), Logistic Regression, Decision Tree] were trained on the same biomarker set for benchmarking.

RESULTS

Automated segmentation showed high agreement with reference masks [Dice at ED: right ventricle [RV] cavity 0.984 ± 0.004, left ventricle (LV)] myocardium 0.965 ± 0.009, LV cavity 0.989 ± 0.003; ES: RV cavity 0.979 ± 0.013, LV myocardium 0.975 ± 0.009, LV cavity 0.985 ± 0.005). The hierarchical veto-logic strategy (V3) achieved an accuracy of 0.925 (95% confidence interval: 0.863–0.975) and a macro-F1 of 0.924, remaining stable across temperatures, outperforming V2 (accuracy 0.787–0.800) and V1 (0.562–0.600). Reproducibility was highest for V3 at T = 0.1 (Fleiss' kappa: 0.969) with a low failure rate (0.83%). Narrative generation produced 97.5% valid reports with 100% numeric fidelity and audited safety ≥ 97.5%. Performance was comparable to supervised models (Random Forest accuracy 0.938; SVM/Logistic Regression accuracy 0.925).

CONCLUSION

In this single-dataset internal evaluation, a glass-box workflow combining automated segmentation-derived biomarkers with an LLM enables robust multiclass cardiac MRI diagnosis while producing numerically faithful, safety-audited narratives, supporting auditability and governance for radiology artificial intelligence (AI). External multicenter validation is needed to confirm generalizability.

CLINICAL SIGNIFICANCE

A glass-box, biomarker-driven agentic-style workflow enables auditable cine cardiac MRI classification with numerically grounded explanations, addressing interpretability and stability barriers that limit translation of radiology AI into routine practice.

KEYWORDS

Cardiac magnetic resonance imaging, explainable artificial intelligence, large language model, quantitative biomarkers

Corresponding author: Ismail Mese

E-mail: ismail_mese@yahoo.com

Received 25 March 2026; revision requested 16 April 2026; accepted 24 April 2026.

Epub: 11.05.2026

Publication date:

DOI: 10.4274/dir.2026.264016

You may cite this article as: Mese I, Kocak B. Glass-box agentic-style workflow for multiclass cine cardiac magnetic resonance imaging classification with a large language model. *Diagn Interv Radiol*. 11 May 2026 DOI: 10.4274/dir.2026.264016 [Epub Ahead of Print].

Artificial intelligence (AI) is increasingly integrated into radiology to support image interpretation, prioritization, and workflow efficiency. One of the earliest clinical implementations was computer-aided detection in mammography in the 1980s.[1] Since then, radiology AI research has expanded rapidly, and commercial translation has accelerated.[2] Reported diagnostic performance has also improved in recent years. Despite these advances, evidence for external validity, clinical impact, and safety in routine practice remains limited.[3,4] This limitation is especially consequential in high-stakes imaging decisions, where models must operate reliably across heterogeneous scanners, acquisition protocols, and patient populations.

A major driver of recent performance gains has been the increasing complexity of models, particularly deep learning methods that learn hierarchical image representations from large datasets for detection and classification tasks.[5] However, performance gains often come at the expense of transparency. Modern models can be difficult to interpret, their failure modes may be hard to anticipate, and their outputs are not always directly linked to clinically recognizable measurements. This performance–explainability trade-off raises concerns regarding auditability, accountability, and governance.[5]

In parallel, AI is evolving from image classification algorithms to large language models (LLMs); LLMs can generate text, summarize longitudinal context, and support reporting.[6-8] More recently, "agentic AI" has emerged as a further step; rather than a single model producing a static output, agentic systems orchestrate multiple agents and modules to complete multistep tasks with limited human supervision.[9] In radiology, such systems could coordinate workflow-level activities, such as protocol selection, automated triage, the synthesis of imaging findings with clinical context, and the drafting of structured and narrative reports for radiologist review.[9] Yet, these capabilities amplify existing safety concerns. Simply put, generated text may be insufficiently constrained by quantitative imaging evidence. Additionally, outputs can vary across repeated runs or decoding settings.[10,11]

Current radiology AI applications, therefore, tend to fall into two categories. End-to-end deep learning systems can achieve high diagnostic accuracy but may offer limited interpretability and limited support for audit.[12,13] On the other hand, LLM-based reporting tools can generate fluent narratives, but these narratives may be weakly grounded in imaging measurements and sensitive to stochastic variability.[10,11] Moreover, clinical radiology lacks widely deployed autonomous and end-to-end agentic systems, and there is limited evidence on how to design workflows that are simultaneously accurate, auditable, and robust to stochastic generation.

To address this, we present a glass-box, agentic-style pipeline for diagnostic decision support that separates perception from reasoning. We use the term "agentic-style" to denote a modular, multistage architecture in which functionally distinct components operate sequentially on structured intermediate representations, with each module receiving explicit inputs and producing auditable outputs. This task decomposition parallels the modular orchestration characteristic of agentic AI systems, in which specialized agents are coordinated to complete complex multistep workflows. The pipeline implements a predetermined, author-designed protocol under explicit human oversight. In this study, cardiac magnetic resonance imaging (MRI) is automatically segmented, transformed into a predefined panel of interpretable biomarkers, and processed through an auditable decision protocol executed by an LLM, which also generates a constrained narrative explanation. We evaluate diagnostic accuracy against standard machine learning baselines, quantify stability across repeated runs and decoding temperatures, and assess whether explanations remain faithful to the underlying quantitative evidence.

Main points

- A glass-box agentic-style workflow separates perception from reasoning: cine cardiac magnetic resonance imaging segmentation is translated into a fixed panel of 17 interpretable biomarkers for protocolized large language model decision-making.
- Baseline prompting (V1) yielded lowest accuracy (0.562–0.600) and showed greater variability across decoding temperatures (T = 0.1–2.0).
- Structured prompting (V2) improved performance to an accuracy range between 0.787 and 0.800, with better temperature robustness across T = 0.1–2.0 than V1.
- Hierarchical, rule-constrained prompting (V3) achieved high diagnostic accuracy (0.925) and remained stable across decoding temperatures (T = 0.1–2.0).
- A decoupled narrative module produced radiologist-style reports with 100% numeric fidelity and audited clinical safety exceeding 95.0%.

Methods

To simulate an agentic-style workflow, a two-stage architecture comprising a perception module and a reasoning core was designed (Figure 1). The study was not integrated into a live clinical workflow, and model outputs did not influence patient care decisions. This study is a secondary analysis of a publicly available anonymized dataset; therefore, additional institutional review board approval and informed consent were not required. Reporting was structured in accordance with the Reporting Checklist for Foundation and Large Language Models (REFINE).[14,15]

Dataset

The Automated Cardiac Diagnosis Challenge (ACDC) cine cardiac MRI dataset was used.[16] The dataset was released in 2017 as part of the ACDC held in conjunction with the Medical Image Computing and Computer-Assisted Intervention 2017 conference. The dataset comprises 150 patients (100 training, 50 test) from a single-center cohort of real clinical data across five diagnostic groups: dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), myocardial infarction with reduced left ventricular ejection fraction (MINF), abnormal right ventricle (RV), and normal controls. Short-axis cine bSSFP images spanning the left ventricle (LV) from base to apex were acquired over the cardiac cycle (two-dimensional slices over time) using breath-hold electrocardiography gating on two Siemens scanners (Avanto 1.5T and Trio Tim 3.0T; Siemens Healthineers, Erlangen, Germany). Acquisition parameters included 5–10-mm slice thickness, 1.34–1.68-mm in-plane resolution, and 28–40 frames per cardiac cycle. Manual reference segmentations of the LV cavity, RV cavity, and LV myocardium at end-diastole (ED) and end-systole (ES), along with diagnostic labels, are provided for the training set. Because the official test labels are withheld, all analyses were performed on the labeled training cohort (n = 100) only. Figure 2 shows representative mid-ventricular slices from the dataset for each class.

Perception module and biomarker extraction

The perception module transformed raw images into a structured biomarker table that served as quantitative input to the downstream reasoning core. This biomarker extraction pipeline utilized a pre-trained nnU-Net model (Task027) to generate automated segmentations of the LV cavity, RV cavity, and LV myocardium at ED and ES.[17] Automated segmentations were evaluated against manual reference masks using the Dice similarity coefficient for each structure at both cardiac phases.

A custom feature extraction engine implemented in Python (NiBabel and SciPy) produced a structured biomarker set comprising 17 quantitative indices (Table 1). The biomarker set was intentionally restricted to clinically interpretable biomarkers, ensuring that downstream reasoning remained grounded in physiologically meaningful measurements rather than latent, black-box representations. Guided by the Society for Cardiovascular Magnetic Resonance reporting recommendations,[18] the set included standard volumetric and functional indices [ED and ES volumes, stroke volumes, and ejection fractions (EFs) for both chambers], as well as myocardial mass and maximal ED wall thickness.

To enhance discrimination of the phenotypes defined by the ACDC, standard metrics were augmented with derived composite markers, including the RV-to-LV ED volume ratio (RV dominance), LV sphericity index, wall thickness variance, systolic myocardial thickening percentage, myocardial mass-to-volume ratio, and the inter-ventricular EF difference [left ventricular ejection fraction (LVEF-RVEF)]. Exact mathematical definitions for all features are provided in Supplementary Material 1.

Automated quality-control rules were applied to filter physiologically implausible values (e.g., EF > 100.0%) prior to analysis.

Reasoning core

The reasoning core was implemented using GPT-OSS-120B (OpenAI, San Francisco, CA), an open-weight LLM released under the Apache 2.0 license.[19] The system is a 117-billion-parameter mixture-of-experts transformer optimized for agentic workflows;[20] GPT-OSS-120B, released on August 5, 2025, is a text-only LLM trained primarily on English-language data. This model was selected for two reasons: (1) scientific reproducibility enabled by transparent weight access and (2) reasoning performance supported by the "Harmony" response format, which facilitates structured inference aligned with our protocol. Inference was executed via the Groq (Groq Inc., Mountain View, CA, USA) hardware-acceleration application programming interface (API) to enable low-latency deployment and consistent runtime behavior across experiments. The model's documented knowledge cut-off is August 2025.

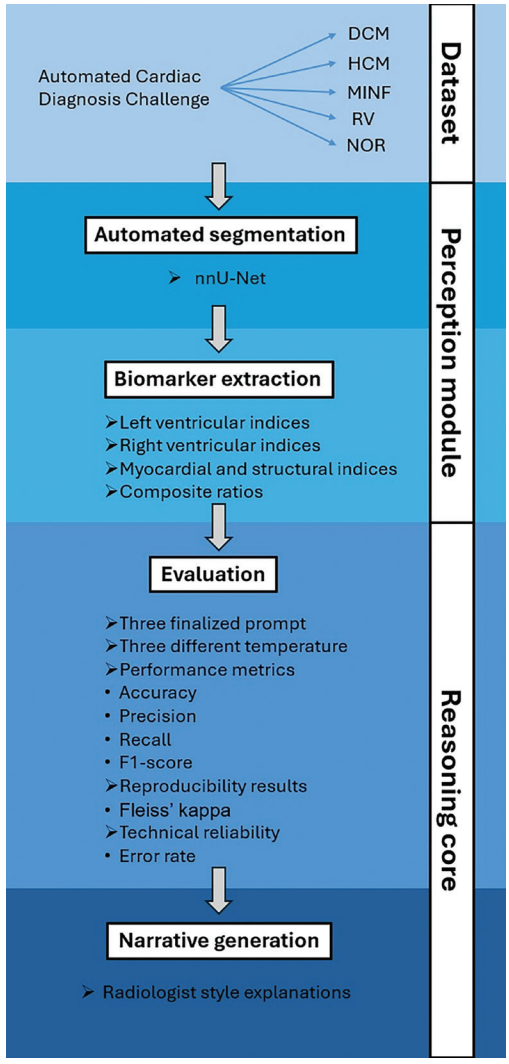

Figure 1. Glass-box radiology pipeline. DCM, dilated cardiomyopathy; HCM, hypertrophic cardiomyopathy; MINF, myocardial infarction with reduced left ventricular ejection fraction; NOR, normal control; RV, abnormal right ventricle.

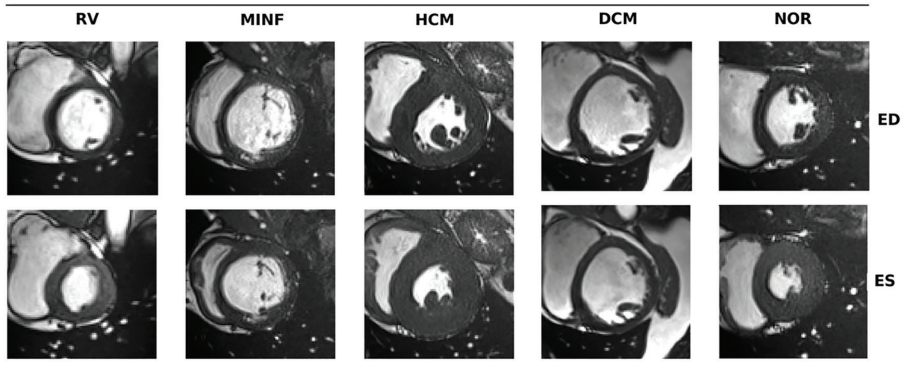

Figure 2. Representative midventricular slices from the Automated Cardiac Diagnosis Challenge dataset. RV, right ventricle; MINF, myocardial infarction with reduced left ventricular ejection fraction; HCM, hypertrophic cardiomyopathy; DCM, dilated cardiomyopathy; NOR, normal control; ED, end-diastole; ES, end-systole.

The model was used in an inference-only configuration without fine-tuning, post-training weight updates, retrieval augmentation, or external tool use.

Prompt engineering

A random stratified split by diagnostic class into a prompt-development set (n = 20) and an independent evaluation set (n = 80) was performed. Prompt optimization and rule finalization were performed exclusively on the prompt-development set; the evaluation set was not accessed during protocol design. All prompts were developed and administered by a radiologist with 3 years of experience in the application of LLMs and were queried in English and in text-only format.

An iterative prompt-engineering phase was implemented, focused on reducing hallucination and enforcing clinically grounded decision-making. A Chain-of-Thought paradigm was used during development to improve intermediate reasoning; however, at inference time, the model output was constrained to a single classification label in JSON format to ensure reliable parsing.

To mitigate failure cases, vague clinical descriptors were replaced with explicit quantitative constraints derived from the prompt-development cohort. Where possible, decision thresholds were informed by established clinical guidelines.[18,21,22] For example, maximal wall thickness ≥ 15.5 mm as the primary criterion for HCM was adopted from the European Society of Cardiology guideline,[21] and LVEF cutoffs for grading systolic dysfunction were informed by current heart failure classification criteria.[22] However, some thresholds used in the V3 protocol, particularly those governing absolute volumetric cutoffs for chamber dilation and composite biomarkers such as RV-to-LV volume ratio, myocardial mass-to-volume ratio, and inter-ventricular EF difference, were empirically refined on the 20-patient prompt-development cohort to optimize separability among the five ACDC phenotypes. In this refinement, for example, the ambiguous instruction "assess for dilation" was replaced with "Check LVESV > 130 mL" because the former frequently led to false-positive dilation in borderline-normal cases. Stop-logic commands were also introduced to prevent drifting into lower-probability diagnoses after a dominant phenotype was established. These analyses resulted in three finalized prompts: Prompt V1 (intrinsic knowledge), Prompt V2 (static constraint), and Prompt V3 (hierarchical veto logic). Full prompt texts are provided in Supplementary Material 2.

Table 1. Description and clinical relevance of the quantitative biomarker panel

Biomarker symbol	Description	Clinical relevance
LVEDV	LV end-diastolic volume (mL)	Primary index of LV dilation.
LVESV	LV end-systolic volume (mL)	Marker of systolic residual volume.
LVSV	LV stroke volume (mL)	Measure of forward output.
LVEF	LV ejection fraction %	Key index of global systolic function.
RVEDV	RV end-diastolic volume (mL)	Index of RV dilation.
RVESV	RV end-systolic volume (mL)	Marker of RV systolic burden.
RVSV	RV stroke volume (mL)	Measure of RV output.
RVEF	RV ejection fraction %	Key index of RV systolic function.
MYO_mass_ED	Myocardial mass at ED (g)	Volumetric proxy for hypertrophy.
MYO_mass_ES	Myocardial mass at ES (g)	Consistency check for mass conservation.
Max_Wall_Thickness	Maximal wall thickness (mm)	Primary diagnostic criterion for HCM.
Sphericity_Index	LV sphericity index	Geometric remodeling marker.
Wall_Variance	LV wall thickness variance	Marker of asymmetric hypertrophy or regional scarring.
Systolic_Thickening	Systolic wall thickening (%)	Index of regional contractile vigor.
RV_LV_ED_ratio	RVEDV/LVEDV	Quantifies RV vs. LV dominance.
MYO_LV_ED_ratio	MYO_mass_ED/LVEDV	Captures hypertrophy relative to cavity size.
EF_Difference	Difference: LVEF-RVEF	Differentiates isolated vs. biventricular failure.

HCM, hypertrophic cardiomyopathy, LVEDV, left ventricular end-diastolic volume; LVESV, left ventricular end-systolic volume; LVSV, left ventricular stroke volume; LVEF, left ventricular ejection fraction; RVEDV, right ventricular end-diastolic volume; RVESV, right ventricular end-systolic volume; RVSV, right ventricular stroke volume; RVEF, right ventricular ejection fraction; MYO, myocardial mass; ED, end-diastole; ES, end-systole; RV, right ventricular; LV, left ventricular; EF, ejection fraction.

Finally, to decouple diagnostic inference from language generation, a dedicated narrative generator prompt was implemented, which synthesized a radiologist-style explanation conditioned on the fixed consensus label and the biomarker table.

Agentic-style implementation

For each patient case in the independent evaluation set (n = 80), raw data were preprocessed to prevent data leakage and memorization artifacts, including the removal of patient identifiers and diagnostic labels.

To mitigate stochastic variability, a majority-vote self-consistency protocol was used. For each case, the diagnostic agent was queried three times via independent API calls using the same prompt under identical decoding settings, with no persistent memory, retained session history, or cross-case context carryover. Each response was restricted to a single label. Outputs were aggregated by majority vote across valid, parsable labels. Unparsable outputs were counted as errors and excluded from the voting pool. In the event of a tie among valid outputs, the first successfully parsed label was used as a deterministic tie-breaker.

To assess stability, a systematic temperature parameter variation analysis (T = 0.1, 1.0, 2.0) was performed, with nucleus sampling fixed (top-p = 1.0) and top-k disabled (top-k = 0). The maximum output length was capped at 1,024 tokens with a "low" reasoning-effort setting to encourage direct classification without verbose reasoning text.

Narrative synthesis with reliability check

After establishing a consensus diagnosis, the pipeline triggered a secondary narrative generator module (T = 0.3), conditioned on the fixed consensus label and the biomarker table. This explicitly decoupled diagnostic accuracy from linguistic quality. For narrative synthesis, the maximum output length was 900 tokens with "medium" reasoning effort. Consensus labels were evaluated against the ground-truth diagnostic labels available in the ACDC training cohort.

To quantify the reliability of the generated text, the same radiologist who performed prompt administration conducted a structured audit of the narratives using a predefined checklist comprising three domains: numeric fidelity, clinical safety, and structural compliance. Numeric fidelity was assessed by cross-referencing each quantitative claim (including ventricular volumes, EFs, and wall thickness values) against the source biomarker data table on a per-value basis; a narrative was considered numerically faithful

only if all reported values exactly matched the biomarker table. Clinical safety was evaluated by systematically screening for contradictions between qualitative descriptors and quantitative biomarkers, specifically targeting errors such as characterizing severe systolic dysfunction as "preserved," mislabeling normal ventricular dimensions as "markedly dilated," or describing reduced EF as "within normal limits." Each narrative was scored on a binary pass/fail basis for clinical safety, and the overall safety score was calculated as the proportion of audited valid narratives that passed all safety checklist items. Structural compliance was assessed by verifying adherence to the prescribed reporting format, defined as single-paragraph radiologist-style prose without bulleting, stepwise enumeration, or extraneous formatting. An additional typographical check confirmed that no narrative contained the prompt instructions themselves or system-level artifacts in the output text.

Baseline machine learning models

To benchmark the proposed framework against supervised learning, Random Forest, Support Vector Machine (SVM), Logistic Regression, and Decision Tree classifiers were implemented using scikit-learn (v1.3.0). All models used the same 17-feature input.

Baseline classifiers were evaluated in two complementary ways. First, to mirror the agent development workflow, hyperparameters were tuned using 5-fold cross-validation (CV) on the prompt-development cohort (n = 20). The best-performing configuration was then refit on the full development cohort and evaluated once on an independent evaluation set (n = 80). Second, as an internal sensitivity analysis, nested stratified CV was performed across the full dataset (n = 100), using 5 outer folds for performance estimation and 5 inner folds for hyperparameter selection.

Statistical analysis

Analyses were performed in Python (v3.10) using scikit-learn (v1.3.0) and statsmodels. Continuous variables were reported as mean ± standard deviation (SD) or median [interquartile range (IQR)], depending on normality; categorical variables were reported as counts and percentages. Diagnostic performance was quantified using accuracy and macro-averaged precision, recall, and F1-score; 95% confidence intervals were computed using non-parametric bootstrapping with 1,000 resamples. In baseline comparison, held-out evaluation results are reported with 95% confidence intervals; nested CV results are reported as mean ± SD across outer folds.

Inter-run agreement across the three self-consistency repetitions was quantified using Fleiss' kappa. Kappa values were interpreted as follows: < 0 = no agreement; 0–0.20 = slight; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = substantial; 0.81–1.00 = almost perfect.[23] Technical reliability was assessed using the error rate, defined as the proportion of unparsable inference calls over the total query volume.

Results

Cohort characteristics and perception results

The cohort (n = 100) was distributed evenly across five diagnostic categories. No patients were excluded after automated quality-control checks of derived biomarkers. Automated nnU-Net segmentations demonstrated high agreement with expert annotations across all cardiac structures at both phases. Mean Dice similarity coefficients at ED were 0.984 ± 0.004 for the RV cavity, 0.965 ± 0.009 for the LV myocardium, and 0.989 ± 0.003 for the LV cavity. At ES, mean Dice coefficients were 0.979 ± 0.013 for the RV cavity, 0.975 ± 0.009 for the LV myocardium, and 0.985 ± 0.005 for the LV cavity. Figure 3 shows representative raw images and corresponding segmentation outputs.

Quantitative phenotypes were consistent with expected remodeling patterns. DCM demonstrated marked LV dilation (mean LVEDV: 284.6 ± 47.8 mL) and severe systolic dysfunction [median LVEF: 16.0% (IQR: 13.1–23.8)]. In contrast, HCM showed pronounced myocardial hypertrophy (mean maximal wall thickness: 21.8 ± 3.7 mm) with preserved systolic function (mean LVEF: 67.4 ± 8.9). Class-stratified biomarker summaries are provided in Table 2.

Model performance across prompt strategies and temperature

The unstructured baseline prompt (V1) demonstrated modest diagnostic performance, with accuracy ranging from 0.562 to 0.600 across temperatures and macro-averaged F1-scores ranging from 0.518 to 0.551. Incorporating static constraints (V2) substantially improved performance, with accuracies between 0.787 and 0.800 and macro-F1 between 0.751 and 0.763.

The hierarchical veto-logic strategy (V3) achieved the highest overall performance and remained stable across all tested temperatures. Accuracy was 0.925 at T = 0.1, 1.0, and 2.0, with macro-precision 0.938, macro-recall 0.925, and macro-F1 0.924 at each temperature. Full metrics are summarized in Table 3.

Per-class evaluation of V3 showed the strongest separability for HCM (precision, recall, and F1-score all 1.000), consistent with a distinct hypertrophy-dominant phenotype. NORs showed perfect precision (1.000) with mildly reduced recall (0.875). Class-wise metrics are shown in Table 4.

Reproducibility and failure analysis

Increasing temperature reduced inter-run agreement and increased failure rates across strategies; V3 demonstrated the highest agreement at T = 0.1 (Fleiss' kappa: 0.969), with a low failure rate of 0.83% (2/240).

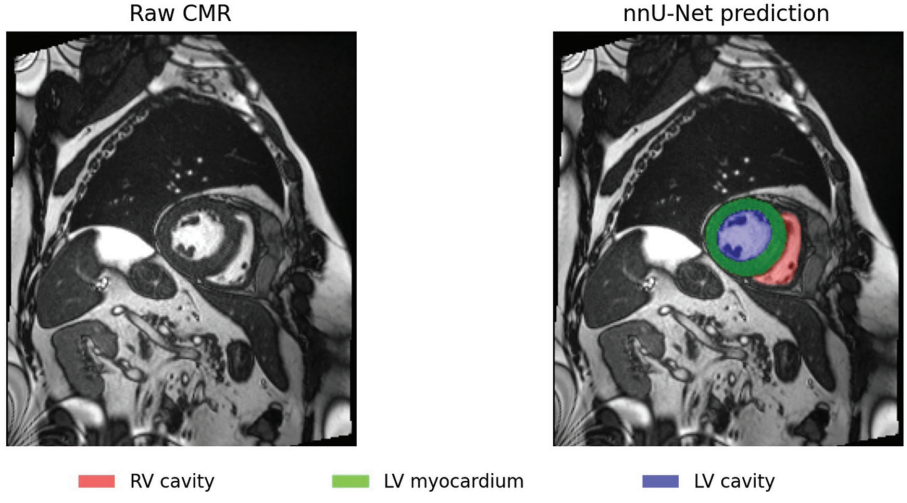

Figure 3. Automated segmentation mask generated by the nnU-Net. CMR, cardiac magnetic resonance imaging; RV, right ventricle; LV, left ventricle.

Table 2. Quantitative biomarker characteristics by diagnostic class (n = 100)

Biomarker	DCM (n = 20)	HCM (n = 20)	MINF (n = 20)	NOR (n = 20)	RV (n = 20)
LVEDV (mL)	284.6 ± 47.8	142.6 (103.0–150.6)	172.2 ± 42.7	130.1 ± 26.4	107.1 ± 37.9
LVESV (mL)	233.1 ± 49.0	42.2 ± 17.6	119.3 ± 35.9	51.7 ±12.4	49.0 ± 20.5
LVSV (mL)	47.6 (37.9–67.5)	86.9 ± 26.8	52.9 ± 18.4	78.4 ± 16.8	58.1 ± 19.1
LVEF (%)	16 (13.1–23.8)	67 ± 9	31 ± 8	60 ± 5	55 ± 6
RVEDV (mL)	178.0 ± 67.4	118.3 ± 34.2	119.1 ± 36.2	153.2 ± 36.6	196.3 ± 48.8
RVESV (mL)	125.6 ± 63.8	46.1 ± 13.4	54.7 ± 23.5	68.8 ± 20.0	134.1 ± 45.1
RVSV (mL)	52.4 ± 27.6	72.2 ± 27.4	64.4 ± 18.5	84.4 ± 25.2	62.2 ± 34.9
RVEF (%)	32 ± 18	60 ± 11	56 ± 10.5	55 ± 8	31 ± 15
MYO_mass_ED (g)	170.2 ± 32.2	177.0 ± 55.1	123.0 ± 18.7	102.4 ± 25.8	77.2 ± 25.5
MYO_mass_ES (g)	183.7 ± 38.6	205.5 ± 59.0	140.0 ± 23.5	119.6 ± 30.3	88.2 ± 29.3
Max_Wall_thickness (mm)	12.8 ± 1.2	21.8 ± 3.7	14.5 ± 2.5	12.5 (11.5–13.7)	12.0 ± 1.6
Sphericity index	0.30 (0.27–0.34)	0.17 (0.15–0.20)	0.31 ± 0.11	0.27 ± 0.09	0.29 ± 0.06
Wall_variance (mm^2)	2.29 ± 0.29	3.85 ± 0.63	2.51 ± 0.40	2.15 ± 0.31	1.97 ± 0.42
Systolic thickening (%)	12.2 ± 7.7	45.1 ± 12.2	27.6 ± 6.9	42.5 ± 6.1	35.9 (26.0–43.3)
RV_LV_ED_ratio	0.62 ± 0.19	0.87 (0.80–0.97)	0.72 ± 0.22	1.17 ± 0.11	1.78 (1.47–2.45)
MYO_LV_ED_ratio	0.60 ± 0.10	1.40 ± 0.33	0.74 ± 0.15	0.79 ± 0.11	0.71 (0.65–0.79)
EF_Difference	−13 ± 15	8 ± 10	−24 ± 14	5 ± 8	23 ± 17

Values are presented as mean ± SD for normally distributed data, and median (IQR) for non-normally distributed data. SD, standard deviation; IQR, interquartile range; DCM, dilated cardiomyopathy; HCM, hypertrophic cardiomyopathy; MINF, myocardial infarction; NOR, normal control; RV, abnormal right ventricle; LVEDV, left ventricular end-diastolic volume; LVESV, left ventricular end-systolic volume; LVSV, left ventricular stroke volume; LVEF, left ventricular ejection fraction; RVEDV, right ventricular end-diastolic volume; RVESV, right ventricular end-systolic volume; RVSV, right ventricular stroke volume; RVEF, right ventricular ejection fraction; MYO, myocardial mass; ED, end-diastole; ES, end-systole; LV, left ventricular; EF, ejection fraction.

Table 3. Performance summary across temperatures for V1, V2, and V3 prompting strategies

Prompt	Accuracy	Precision	Recall	F1-score
V1 (T = 0.1)	0.588 (0.475–0.700)	0.621 (0.516–0.725)	0.588 (0.475–0.700)	0.546 (0.435–0.657)
V1 (T = 1.0)	0.562 (0.450–0.663)	0.586 (0.474–0.697)	0.562 (0.450–0.663)	0.518 (0.397–0.627)
V1 (T = 2.0)	0.600 (0.487–0.713)	0.586 (0.482–0.706)	0.600 (0.487–0.713)	0.551 (0.434–0.667)
V2 (T = 0.1)	0.787 (0.700–0.863)	0.806 (0.647–0.914)	0.787 (0.700–0.863)	0.751 (0.650–0.848)
V2 (T = 1.0)	0.800 (0.713–0.887)	0.822 (0.658–0.927)	0.800 (0.713–0.887)	0.763 (0.666–0.865)
V2 (T = 2.0)	0.800 (0.713–0.875)	0.822 (0.660–0.926)	0.800 (0.713–0.875)	0.763 (0.662–0.854)
V3 (T = 0.1)	0.925 (0.863–0.975)	0.938 (0.903–0.978)	0.925 (0.863–0.975)	0.924 (0.861–0.975)
V3 (T = 1.0)	0.925 (0.863–0.975)	0.938 (0.903–0.978)	0.925 (0.863–0.975)	0.924 (0.857–0.975)
V3 (T = 2.0)	0.925 (0.863–0.975)	0.938 (0.902–0.978)	0.925 (0.863–0.975)	0.924 (0.859–0.975)

T = temperature.

Agreement declined at higher temperatures (kappa: 0.917 at T = 1.0; 0.877 at T = 2.0), along with increased failure rates of 3.33% and 5.42%, respectively; V1 showed the greatest degradation at high temperature (kappa: 0.676 at T = 2.0). Full reproducibility metrics are summarized in Table 5.

Table 4. Per-class precision, recall, and F1-score for V3 at T = 0.1

Class	Precision	Recall	F1-score	Support
DCM	0.800	1.000	0.889	16
HCM	1.000	1.000	1.000	16
MINF	1.000	0.750	0.857	16
NOR	1.000	0.875	0.933	16
RV	0.889	1.000	0.941	16

DCM, dilated cardiomyopathy; HCM, hypertrophic cardiomyopathy; MINF, myocardial infarction; NOR, normal control; RV, right ventricle; T, temperature.

Glass-box radiology

The narrative reports generated by GPT-OSS-120B were analyzed across nine experimental runs (n = 80 cases per run). Across all runs, the system produced 702 out of 720 valid narratives (97.5%), with per-run completion rates ranging from 92.5% to 100%. Narratives were compact and consistent in length (mean: 186–211 words per run) and adhered tightly to the required report style, as 0% contained bulleting or stepwise enumeration and outputs exhibited ≤ 1 line break (i.e., a single coherent paragraph with occasional formatting separation). The prompt's intended "radiologist-style" voice was reflected in the outputs. Numeric fidelity was 100% across all valid narratives, and clinical safety audits revealed a safety score

exceeding 97.5%. Narrative reporting details are provided in Supplementary Table 1.

Model-generated narratives for each case in V3 (temperature = 0.1) are accessible via the following OSF reference.[15]

Benchmarking against supervised machine learning

When benchmarked against supervised classifiers trained on the same 17-feature input using the predefined development/evaluation split (n = 20 for tuning/training; n = 80 for held-out evaluation), the V3 strategy demonstrated performance parity with supervised baselines. On the independent evaluation set, Random Forest achieved an accuracy of 0.938 (0.88–0.99), whereas SVM and Logistic Regression achieved 0.925. A single Decision Tree showed lower performance [accuracy: 0.825 (0.74–0.91)]. Figure 4 presents the confusion matrices on the held-out set for the three prompting strategies and the three selected supervised baseline models. Figure 5 shows subgroup-specific F1-score patterns across the held-out evaluations.

As an internal validation sensitivity analysis using nested 5 × 5 stratified CV across the full cohort (n = 100), supervised performance was consistent (e.g., Logistic Regression: 0.950 ± 0.035; Random Forest: 0.940 ± 0.042; SVM: 0.930 ± 0.067; Decision Tree: 0.900 ± 0.050, mean ± SD across outer folds). Comparative metrics are shown in Table 6.

Table 5. Reproducibility and failure analysis across decoding temperatures for V1, V2, and V3

Prompt	Fleiss' kappa	Failure count	Failure rate
V1 (T = 0.1)	0.874	0/240	0.00%
V1 (T = 1.0)	0.852	2/240	0.83%
V1 (T = 2.0)	0.676	10/240	4.17%
V2 (T = 0.1)	0.881	9/240	3.75%
V2 (T = 1.0)	0.880	5/240	2.08%
V2 (T = 2.0)	0.839	14/240	5.83%
V3 (T = 0.1)	0.969	2/240	0.83%
V3 (T = 1.0)	0.917	8/240	3.33%
V3 (T = 2.0)	0.877	13/240	5.42%

T = temperature.

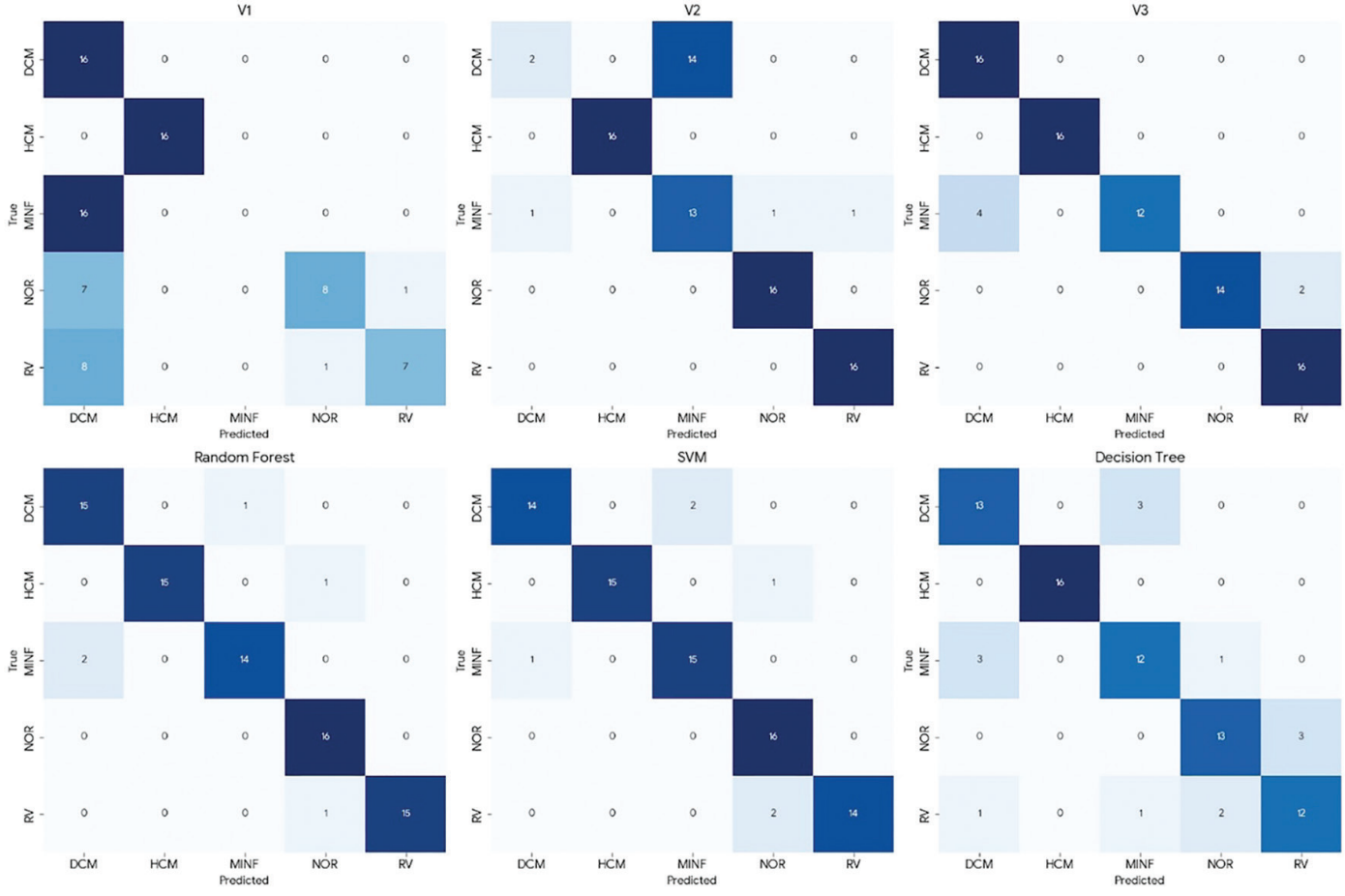

Figure 4. Confusion matrices for the three prompt strategies and three selected baseline models (held-out evaluation). RV, right ventricle; MINF, myocardial infarction with reduced left ventricular ejection fraction; HCM, hypertrophic cardiomyopathy; DCM, dilated cardiomyopathy; NOR, normal control; ED, end-diastole; ES, end-systole; SVM, Support Vector Machine.

Discussion

In this study, we developed a glass-box, agentic-style radiology pipeline that separates perception from reasoning and enables auditable decision-making. In the ACDC cine cardiac MRI cohort, the hierarchical, rule-constrained prompting strategy (V3) achieved high diagnostic accuracy (0.925) and remained stable across decoding temperatures (T = 0.1–2.0). The decoupled narrative generator produced radiologist-style reports with complete numeric fidelity and high audited safety. Performance was comparable to supervised machine learning baselines trained on the same biomarker set.

Taken together, these findings are clinically relevant given that many AI tools for radiology image interpretation remain difficult to audit. End-to-end machine learning and deep learning systems typically rely on latent representations that are not directly traceable to clinically recognizable measurements, limiting interpretability, error analysis, and governance.[24,25] By contrast, our glass-box pipeline constrains inference to a predefined biomarker panel and an explicit decision protocol, providing a reviewable chain from image-derived measurements to the final label. In addition, the decoupled narrative generator translates the same quantitative evidence into a radiologist-style report, providing a human-readable explanation that can be checked directly against the biomarker table. This level of transparency supports reproducibility and quality assurance and may help translate research performance into safe deployment in routine clinical practice.

Automated nnU-Net segmentation yielded high Dice similarity for ventricular cavities and myocardium, indicating reliable extraction of quantitative inputs for downstream reasoning. This supports the use of a compact biomarker table as a stable interface between cine imaging and decision logic; remaining errors likely reflect phenotype overlap and threshold sensitivity rather than segmentation failure.

Baseline prompt (V1) performance was modest, indicating that the model's internal priors alone were insufficient for reliable phenotype assignment. Imposing an explicit, quantitatively grounded protocol (V2) substantially improved accuracy and macro-F1, with further gains under the hierarchical veto-logic scheme (V3) and preserved performance across sampling temperatures. This supports constraining the model to execute transparent decision rules over interpretable biomarkers rather than relying on unconstrained diagnostic heuristics.

Increasing temperature predominantly reduced reproducibility and output parsability. Although majority-vote accuracy for V3 remained nominally similar across all tested

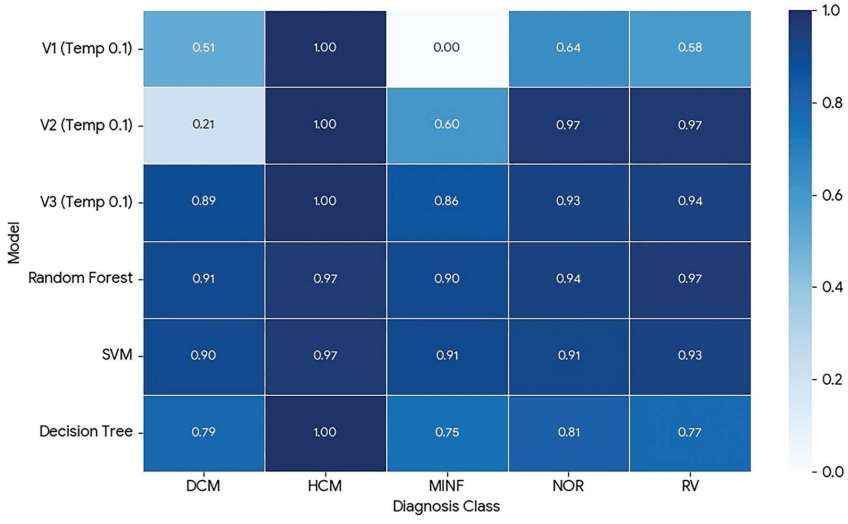

Figure 5. Subgroup performance for prompt strategies and baseline models (F1-score heatmap for held-out evaluations). SVM, Support Vector Machine; DCM, dilated cardiomyopathy; HCM, hypertrophic cardiomyopathy; MINF, myocardial infarction with reduced left ventricular ejection fraction; NOR, normal control; RV, right ventricle.

Table 6. Comparative performance metrics of classification models

Model (evaluation)	Accuracy	Precision	Recall	F1-score	Macro AUC
V3 (T = 0.1) (held-out)	0.925 (0.86–0.98)	0.938 (0.89–0.98)	0.925 (0.87–0.98)	0.924 (0.85–0.98)	—
Random Forest (held-out)	0.938 (0.88–0.99)	0.941 (0.89–0.99)	0.938 (0.88–0.99)	0.938 (0.88–0.99)	—
Random Forest (nested CV)	0.940 ± 0.042	0.948 ± 0.039	0.940 ± 0.042	0.939 ± 0.042	0.996 ± 0.007
SVM (held-out)	0.925 (0.86–0.98)	0.932 (0.88–0.98)	0.925 (0.87–0.98)	0.926 (0.86–0.98)	—
SVM (nested CV)	0.930 ± 0.067	0.943 ± 0.056	0.930 ± 0.067	0.928 ± 0.070	0.996 ± 0.004
Logistic Regression (held-out)	0.925 (0.88–0.98)	0.936 (0.89–0.98)	0.925 (0.87–0.98)	0.926 (0.87–0.98)	—
Logistic Regression (nested CV)	0.950 ± 0.035	0.956 ± 0.036	0.950 ± 0.035	0.950 ± 0.035	0.998 ± 0.003
Decision Tree (held-out)	0.825 (0.74–0.91)	0.825 (0.74–0.91)	0.825 (0.74–0.91)	0.825 (0.74–0.90)	—
Decision Tree (nested CV)	0.900 ± 0.050	0.921 ± 0.040	0.900 ± 0.050	0.899 ± 0.050	0.936 ± 0.031

AUC, area under the curve; CV, cross-validation; T, temperature; SVM, Support Vector Machine.

temperatures, inter-run agreement declined with higher temperature (Fleiss' kappa: from 0.969 at T = 0.1 to 0.877 at T = 2.0), accompanied by increased failure rates due to non-parsable outputs (from 0.83% at T = 0.1 to 5.42% at T = 2.0). Accordingly, the characterization of V3 as "stable" applies specifically to majority-vote accuracy; per-run reproducibility and parsing reliability degraded at higher temperatures. These findings support low-temperature decoding and strictly structured outputs for safety-oriented deployment when stochastic sampling is required.

The hierarchical veto-logic strategy (V3) relies on quantitative thresholds, partly derived from established clinical guidelines[21,22] and partly refined empirically on the 20-patient prompt-development cohort. Although guideline-aligned thresholds are expected to transfer across populations, the empirically refined cutoffs carry an inherent risk of dataset-specific optimization given the small development sample (n = 20). Importantly, the progression from V1 to V3 provides indirect evidence regarding the necessity of explicit constraints; V1, which relied entirely on the model's internal medical knowledge without imposed thresholds, achieved only modest accuracy (0.562–0.600), demonstrating that the LLM's pretrained clinical priors alone are insufficient for reliable phenotype assignment in this task. The substantial performance gain observed with V2 and V3 confirms that structured, quantitatively grounded decision protocols drive diagnostic accuracy in the proposed framework. Nevertheless, because threshold optimization was performed on a limited cohort, external validation across different sites, scanners, field strengths, and patient demographics will be essential to determine whether the current cutoffs generalize or require site-specific recalibration.

An important consideration is the incremental value of the proposed agentic-style framework relative to conventional supervised classifiers or symbolic rule engines. As demonstrated in our benchmarking analysis, supervised models trained on the same 17-feature biomarker set achieved comparable classification accuracy. The V3 prompting strategy, in particular, functions similarly to an expert-designed rule-based classifier expressed in natural language form. We explicitly acknowledge that the primary diagnostic performance of V3 derives from the structured decision logic rather than from emergent model intelligence. However, the critical distinction is that a conventional classifier produces only a label, whereas the agentic-style framework integrates classification and narrative explanation generation within a single auditable pipeline. A Random Forest or SVM achieving 0.938 accuracy still requires a separate, independently maintained reporting system to translate its output into clinically actionable text, and the fidelity between that classifier's decision and the resulting narrative must be verified through additional mechanisms. Furthermore, the decision rules are expressed in natural language that clinicians can directly inspect, critique, and modify without programming expertise, and prompt-based protocols can be adapted to evolving diagnostic criteria or new biomarker panels through text editing rather than model retraining. These operational and governance-related advantages represent the primary value proposition of the agentic-style approach.

A sharper version of this comparison considers a deterministic symbolic rule engine paired with a templated narrative generator, which would replicate V3's classification logic in code and emit fixed reporting strings. The agentic-style framework differs from such a baseline in three respects. The first concerns narrative behavior under case variability. Templated outputs require enumerated branches to handle atypical or borderline biomarker constellations, whereas an LLM narrative conditioned on the same biomarker table adapts phrasing to the specific value pattern without combinatorial template expansion. The second concerns single-source coherence. A rule-plus-template pipeline maintains classification logic and narrative text as separate artifacts that can drift out of sync when thresholds or reporting conventions are updated, whereas in the proposed pipeline, the decision protocol and the narrative generator are both grounded in the same biomarker table and consensus label, making coherence structural rather than procedural. The third concerns extensibility. Adding a new biomarker, phenotype, or non-imaging input to a rule-plus-template system requires coordinated updates across the rule code, template library, and their linking branches, whereas the prompt-based architecture extends only to additions to the biomarker table and prompt text.

A decoupled narrative stage supports a glass-box radiology workflow in which interpretation is derived from an explicit biomarker table and a predefined decision protocol, and the narrative functions as a reporting layer. This contrasts with black-box approaches that produce final labels without accessible intermediate evidence. The separation improves traceability between measurements, decision rules, and reported conclusions, enabling routine audit, targeted quality assurance, and clinician verification. This architecture is advantageous for clinical translation because model updates can be implemented through protocol refinement, threshold recalibration, or incorporation of additional biomarkers, without retraining an opaque end-to-end classifier. Error analysis becomes actionable, as failures can be attributed to specific measurements or decision steps, facilitating systematic remediation while preserving interpretability and governance.

In the ACDC classification challenge, leading submissions achieved accuracies of 0.86–0.96 on the 50-patient test set. Khened et al.[26] reported the highest accuracy (0.96) using a Random Forest trained on 11 predominantly segmentation-derived features, supplemented by patient height and weight. Isensee et al.[27] achieved an accuracy of 0.92 using dynamic and instantaneous features extracted from segmentation maps, and an ensemble combining multiple multilayer perceptrons with a Random Forest. Wolterink et al.[28] reported an accuracy of 0.86 using 14 mainly segmentation-derived features and a larger Random Forest. Our approach is within this performance range (V3 accuracy 0.925) while remaining fully image-driven: automated segmentation is translated into a compact, predefined biomarker panel that directly governs the downstream decision, enabling numerically auditable explanations. Comparable performance has also been reported for report-based LLM approaches. Zaman et al.[29] reported a micro-F1 of 0.86 and a micro-area under the curve of 0.96 for five diagnoses from 1,503 cardiac MRI reports, and Wang et al.[30] reported improved prompted-LLM performance for myocardial infarction/DCM/HCM, with GPT-4 reaching an accuracy of 95.8% under informative prompting and high agreement with radiologists (AC1: approx. 0.93). In contrast to these text-dependent methods, our workflow operates on raw cine images and uses a fixed set of quantitative biomarkers to drive classification.

An important advantage of using an open-weight LLM is the potential for improved transparency, reproducibility, and institutional control compared with fully closed systems. Such models may enable deployment within secure local infrastructure or regionally compliant environments, thereby strengthening data security and privacy protections when sensitive clinical information is processed. Furthermore, pre-

serving inference logs and decision traces may enhance auditability by enabling systematic review of model behavior, output provenance, and potential safety failures. These characteristics may be valuable for governance, accountability, and responsible oversight of clinical AI systems. However, the present framework remains a research-stage diagnostic decision-support tool and should not be interpreted as a validated autonomous diagnostic system.

This study has several limitations. First, the proposed workflow addresses only the cine-derived volumetric and functional component of cardiac MRI interpretation. In routine clinical practice, cardiac MRI diagnosis frequently relies on additional sequences and information beyond cine imaging, particularly late gadolinium enhancement for myocardial scar detection and viability assessment; native T1 and T2 mapping for tissue characterization, including edema, diffuse fibrosis, and infiltrative disease; and integration with clinical history, laboratory data, electrocardiographic findings, and prior imaging. Accordingly, a comprehensive clinical decision-support system would need to incorporate multiparametric imaging inputs and contextual clinical data to address the full diagnostic spectrum encountered in practice. Second, the proposed pipeline should not be interpreted as a fully agentic system in the broader sense of agentic AI, which typically implies some degree of autonomy in planning, tool selection, and adaptive decision-making. In this work, we deliberately prioritized transparency, auditability, and methodological control over autonomy by constraining the LLM to a predefined, author-designed decision pipeline. Although this design improves interpretability and oversight, it may limit the flexibility and generalizability that more autonomous, agentic systems could offer in less-structured clinical settings. Third, the evaluation was performed on a single public cardiac MRI dataset with a modest labeled cohort (n = 100) and five phenotypic classes; generalization to other institutions, vendors, acquisition protocols, and more heterogeneous real-world populations remains unproven. Fourth, all diagnostic analyses used the labeled training cohort because official test labels are withheld; although the prompt-development/evaluation split reduces leakage during prompt design, performance may still be optimistic relative to a truly external test set. Fifth, the pipeline depends on upstream segmentation quality and on the specific biomarker definitions chosen. Although Dice scores were high, segmentation errors could propagate into biomarker errors and downstream misclassification; robustness to lower-quality segmentations and to "out-of-distribution" anatomy was not directly tested. Sixth, the hierarchical protocol (V3) uses thresholds derived during prompt development; these cutoffs may be dataset-specific and could require recalibration across sites, scanners, or patient mix. Seventh, narrative quality and safety were audited by a radiologist, but the audit was still manual and may be subject to reviewer bias; multireader audits and standardized scoring rubrics would strengthen the evidence. Finally, this was a retrospective technical evaluation: no prospective workflow study was performed to measure time savings, changes in reporting consistency, or downstream clinical impact.

In conclusion, we propose a glass-box, agentic-style cardiac MRI cine workflow in which a perception module derives interpretable biomarkers, and a diagnostic agent executes an explicit decision protocol with self-consistency to output an auditable consensus label. The hierarchical veto-logic strategy (V3) achieved high, temperature-stable performance (accuracy: 0.925; macro-F1: 0.924 across T = 0.1–2.0), and a separate reporting agent generated numerically faithful narratives, with audited clinical safety exceeding 95.0%. As this evaluation was conducted on a single public dataset without external validation, these results should be interpreted with appropriate caution. External multicenter validation and prospective multireader assessment are required before clinical deployment.

Acknowledgement

Language of this manuscript was checked and improved by ChatGPT (GPT-5.4). The authors conducted strict supervision when using these tools.

Footnotes

Conflict of interest disclosure

The authors declared no conflicts of interest.

Supplementary Table: https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/c9445cd4-3480-4836-a5a1-2b7c0b8f3495.pdf

Supplementary Materials: https://d2v96fxpocvxx.cloudfront.net/beb8919b-f013-4ea1-b1c8-40332e840fe1/content-images/3520c5af-3e54-4ff4-933c-fa6fd-de33e25.pdf

References

1. Bhandari A. Revolutionizing radiology with artificial intelligence. *Cureus*. 2024;16(10):e72646. [Crossref]
2. van Leeuwen KG, de Rooij M, Schalekamp S, van Ginneken B, Rutten MJCM. How does artificial intelligence in radiology improve efficiency and health outcomes? *Pediatr Radiol*. 2022;52(11):2087-2093. [Crossref]
3. Tariq A, Purkayastha S, Padmanaban GP, et al. Current clinical applications of artificial intelligence in radiology and their best supporting evidence. *J Am Coll Radiol*. 2020;17(11):1371-1381. [Crossref]
4. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol*. 2021;31(6):3797-3804. [Crossref]
5. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy*. 2021;23(1):18. [Crossref]
6. Mese I, Kocak B. ChatGPT as an effective tool for quality evaluation of radiomics research. *Eur Radiol*. 2025;35(4):2030-2042. [Crossref]
7. Mese I, Kocak B. Large language models in methodological quality evaluation of radiomics research based on METRICS: ChatGPT vs NotebookLM vs radiologist. *Eur J Radiol*. 2025;184:111960. [Crossref]
8. Mese I, Kocak B. Evaluating methodological quality in radiomics research using large language models: added value of METRICS-E3 framework. *Eur J Radiol*. 2026;194:112519. [Crossref]
9. Koçak B, Meşe İ. AI agents in radiology: toward autonomous and adaptive intelligence. *Diagn Interv Radiol*. 2025. [Crossref]
10. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*. 2024;310(1):e232756. [Crossref]
11. Nakaura T, Ito R, Ueda D, et al. The impact of large language models on radiology: a guide for radiologists on the latest innovations in AI. *Jpn J Radiol*. 2024;42(7):685-696. [Crossref]
12. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med*. 2022;5(1):48. [Crossref]
13. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56. [Crossref]
14. Mese I, Akinci D'Antonoli T, Bluethgen C, et al. Reporting checklist for foundation and large language models in medical research (REFINE): an international consensus guideline. *Diagn Interv Radiol*. 2026. [Crossref]

15. Model-generated narrative reports for the 80-Case independent evaluation set using GPT-OSS-120B and the V3 hierarchical veto-logic prompt. Published online March 17, 2026. Accessed March 18, 2026. [Crossref]
16. Bernard O, Lalande A, Zotti C, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans Med Imaging*. 2018;37(11):2514-2525. [Crossref]
17. Isensee F, Petersen J, Klein A, et al. nnU-Net: self-adapting framework for U-Net-based medical image segmentation. *arXiv*. 2018. [Crossref]
18. Hundley WG, Bluemke DA, Bogaert J, et al. Society for Cardiovascular Magnetic Resonance (SCMR) guidelines for reporting cardiovascular magnetic resonance examinations. *J Cardiovasc Magn Reson*. 2022;24(1):29. [Crossref]
19. openai/gpt-oss-120b · Hugging Face. August 7, 2025. Accessed March 19, 2026. [Crossref]
20. OpenAI, Agarwal S, Ahmad L, et al. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv*. 2025. [Crossref]
21. Authors/Task Force members; Elliott PM, Anastasakis A, Borger MA, et al. 2014 ESC Guidelines on diagnosis and management of hypertrophic cardiomyopathy: the Task Force for the Diagnosis and Management of Hypertrophic Cardiomyopathy of the European Society of Cardiology (ESC). *Eur Heart J*. 2014;35(39):2733-2779. [Crossref]
22. Heidenreich PA, Bozkurt B, Aguilar D, et al. 2022 AHA/ACC/HFSA Guideline for the management of heart failure: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation*. 2022;145(18):e876-e894. [Crossref]
23. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-282. [Crossref]
24. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: a review of interpretability methods. *Comput Biol Med*. 2022;140:105111. [Crossref]
25. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. *J Imaging*. 2020;6(6):52. [Crossref]
26. Khened M, Alex V, Krishnamurthi G. Densely connected fully convolutional network for short-axis cardiac cine MR image segmentation and heart diagnosis using random forest. In: Pop M, Sermesant M, Jodoin PM, et al., eds. Statistical atlases and computational models of the heart. ACDC and MMWHS Challenges. Springer International Publishing; 2018:140-151. [Crossref]
27. Isensee F, Jaeger PF, Full PM, Wolf I, Engelhardt S, Maier-Hein KH. Automatic Cardiac Disease Assessment on cine-MRI via Time-Series Segmentation and Domain Specific Features. In: Pop M, Sermesant M, Jodoin PM, et al., eds. Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Springer International Publishing; 2018:120-129. [Crossref]
28. Wolterink JM, Leiner T, Viergever MA, Išgum I. Automatic Segmentation and Disease Classification Using Cardiac Cine MR Images. In: Pop M, Sermesant M, Jodoin PM, et al., eds. Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges. Springer International Publishing; 2018:101-110. [Crossref]
29. Zaman S, Petri C, Vimalesvaran K, et al. Automatic diagnosis labeling of cardiovascular MRI by using semisupervised natural language processing of text reports. *Radiol Artif Intell*. 2021;4(1):e210085. [Crossref]
30. Wang L, Peng L, Wan Y, et al. Automated cardiac magnetic resonance interpretation derived from prompted large language models. *Cardiovasc Diagn Ther*. 2025;15(4):726-737. [Crossref]