Enhancing CineMRI Clinical Documentation by Detecting and Correcting Ambiguity with Large Language Models

Guillermo Villanueva Benito[1]
Biomedical Data Science Group
[1]Barcelona Institute for Global Health (ISGlobal)
Barcelona, Spain
e-mail: guillermo.villanueva@isglobal.org

Paula Petrone[1,2]
Digital Health Unit
[2]Barcelona Supercomputing Center (BSC-CNS)
Barcelona, Spain
e-mail: paula.petrone@bsc.es

Matias Calandrelli[3], Martín Descalzo[3], Sandra Pujadas[3], Juan Fernandez[3]
Cardiac Imaging Unit, Cardiology Department
[3]Hospital de la Santa Creu i Santa Pau
Barcelona, Spain
e-mail: {mcalandrelli | mdescalzo | sandrapujadas | juanfmav}@gmail.com

***Abstract*— Cardiac cine Magnetic Resonance Imaging (cineMRI) is the gold standard for assessing left-ventricular wall motion, yet interpretation varies and free-text reports often contain ambiguous terminology. We developed CineScribe, an Artificial Intelligence (AI)-assisted framework that structures diagnostic information, detects ambiguous clinical reports, and generates standardized cineMRI documentation. Using a dataset of 982 cineMRI studies, CineScribe achieved state-of-the-art report structuration performance (F1-score = 0.92). Associated confidence scores effectively identified ambiguous cases (F1-score = 0.82), most of which carried misdiagnosis risk. Generated reports from final review findings were rated complete and accurate in 78% of cases, supporting more consistent and reliable cineMRI documentation.**

Keywords- Cardiac cineMRI; LLMs; Documentation; ambiguity.

I. INTRODUCTION

Despite major advances in recent decades, cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, accounting for more than 17.9 million deaths annually [1]. Many diagnostic decisions in cardiology rely heavily on imaging, and cardiac cine Magnetic Resonance Imaging (cineMRI) is the gold standard for evaluating cardiac structure and function. CineMRI provides a comprehensive assessment of global and regional ventricular motion, primarily determined by the presence and severity of Regional Wall Motion Abnormalities (RWMAs). Yet its interpretation remains inherently variable, even among experienced cardiologists. This intrinsic variability is compounded by the lack of standardized reporting practices which lead to clinical reports being often written using imprecise and inconsistent terminology, increasing the risk of misdiagnosis [2]-[4]. Identifying ambiguous reports that may lead to miscommunication, detecting diagnostically complex cases requiring further expert consensus, and implementing standardized review and documentation processes are essential to ensure accurate diagnosis and appropriate downstream patient management.

The rest of the paper is organized as follows: Section II describes the objectives and provides an overview of the framework. Section III details the methodology, Section IV presents the main results, and Section V concludes with a summary and future research directions.

II. OBJECTIVES

To address the challenges presented in the previous section, we developed and validated an AI-assisted framework designed to facilitate more reliable and standardized cineMRI clinical documentation of left ventricular wall motion. The framework, illustrated in Figure 1, integrates three complementary components:

1) Automatic report structuration to extract and organize diagnostic information from free-text reports.
2) Detection of ambiguous or imprecise reports that might lead to misdiagnosis, thereby enabling targeted expert cineMRI review and consensus.
3) Standardized AI-assisted report generation from structured diagnostic findings to ensure clear, complete, and reproducible clinical documentation.

The model first interprets the original free-text cineMRI report, converts it into a structured representation visualized as a bullseye diagram, and assigns a global confidence score reflecting uncertainty in the extracted interpretation. Subsequently, experts reassess the corresponding cineMRI scans for flagged ambiguous reports and provide the final structured diagnostic bullseye. CineScribe then generates a standardized clinical report using the confirmed structured diagnosis and Left Ventricular Ejection Fraction (LVEF).

III. METHODS

We conducted a retrospective study including 982 cineMRI examinations from the Hospital de la Santa Creu i Sant Pau (Barcelona, Spain). Three board-certified cardiologists, with certification in cardiac Magnetic

Resonance Imaging (MRI), manually annotated the clinical reports, assigning a RWMA label to each myocardial segment [5] based on the findings described in the original text. For multi-expert (x3) annotated cases, report ambiguity was defined as inter-expert disagreement, with ambiguous reports corresponding to those that yielded more than one valid structured interpretation from medical experts.

We developed CineScribe, a fine-tuned lightweight Large Language Model (LLM) based on Llama3 [6], trained for both report-structuration and clinical report-generation tasks. The model's confidence score during the structuration task was validated as a quantitative proxy for report ambiguity. Reports were re-evaluated by expert cardiologists, who reviewed the original cineMRI videos to reassess the corresponding diagnostic findings. CineScribe was quantitatively benchmarked for report-structuration performance and prospectively evaluated on a stressed evaluation dataset composed exclusively of abnormal cases by expert cardiologists for its ability to detect ambiguous reports, flag cases at risk of misdiagnosis that require expert review, and for the clinical quality of its generated reports, following the QUEST framework [7].

IV. RESULTS

CineScribe achieved an F1-score of 0.92 (95% Confidence Interval (CI): 0.89 -- 0.94) in the report structuration task, demonstrating state-of-the-art performance comparable to GPT-5 (F1-score = 0.89; 95% CI: 0.85 -- 0.92). The model's confidence score demonstrated strong discriminative ability for detecting ambiguous reports, achieving an Area Under the Receiver Operating Characteristic Curve (ROC–AUC) of 0.76 on the evaluation dataset.

Subsequent expert cineMRI review showed that 71% of non-ambiguous reports were safe-to-follow, whereas 84% of ambiguous reports carried potential misdiagnosis risk, underscoring CineScribe's utility in flagging diagnostically complex cases that warrant targeted expert review.

Finally, standardized clinical reports generated by CineScribe from structured reviewed findings were rated both complete and accurate in 78% of cases by expert cardiologists.

V. CONCLUSIONS AND FUTURE WORK

CineScribe improves cineMRI interpretation by accurately structuring diagnostic information, flagging ambiguous reports that signal diagnostically complex cases, and producing clear and consistent clinical reports. Our findings suggest that ambiguous report language often arises in clinically challenging cases. From a clinical perspective, this has important implications as ambiguous reports can mask diagnostically difficult cases that would benefit from consensus expert review.

Among the limitations of this work is its reliance on data from a single institution, which may restrict the generalizability of the findings, as well as its motion-focused scope, which represents only a subset of full cineMRI clinical reporting. Future work should further explore human-in-the-loop approaches, such as the one proposed here, to improve report quality and support continued fine-tuning and enhancement of the model's capabilities. Further directions also include the integration of image-derived cineMRI features, as well as prospective deployment and validation in multi-center settings. Finally, continued research is needed to better characterize and address the intrinsic sources of variability in cineMRI assessment, particularly in diagnostically complex cases.

REFERENCES

[1] G. A. Mensah, G. A. Roth, and V. Fuster, "The global burden of cardiovascular diseases and risk factors: 2020 and beyond", Journal of the American College of Cardiology, vol. 74, no. 20, pp. 2529-2532, 2019.

[2] A. B. Rosenkrantz, M. Kiritsy, and S. Kim, "How "consistent" is "consistent"? A clinician-based assessment of the reliability of expressions used by radiologists to communicate diagnostic confidence.", Clinical Radiology, vol. 69, no. 7, pp. 745–749, 2014.

[3] J. M. Bosmans, J. J. Weyler, A. M. De Schepper and P. M. Parizel, "The radiology report as seen by radiologists and referring clinicians: results of the COVER and ROVER surveys.", Radiology, vol. 259, no. 1, pp. 184–195, 2011.

[4] W. Levinson, "Physician-patient communication: a key to malpractice prevention.", JAMA, vol. 272, no. 20, pp. 1619–1620, 1994.

[5] M. D. Cerqueira, et al., "Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart: A statement for healthcare professionals from the Cardiac Imaging Committee of the Council on Clinical Cardiology of the American Heart Association", Circulation, vol. 105, no. 4, pp. 539–542, 2002.

[6] A. Grattafiori, et al., "The llama 3 herd of models.", arXiv preprint arXiv:2407.21783, 2024.

[7] K. K. Y. Ng, I. Matsuba, P. C. Zhang, "RAG in Health Care: A Novel Framework for Improving Communication and Decision-Making by Addressing LLM Limitations.", NEJM AI, vol. 2, no. 1, AIra2400380, 2024.

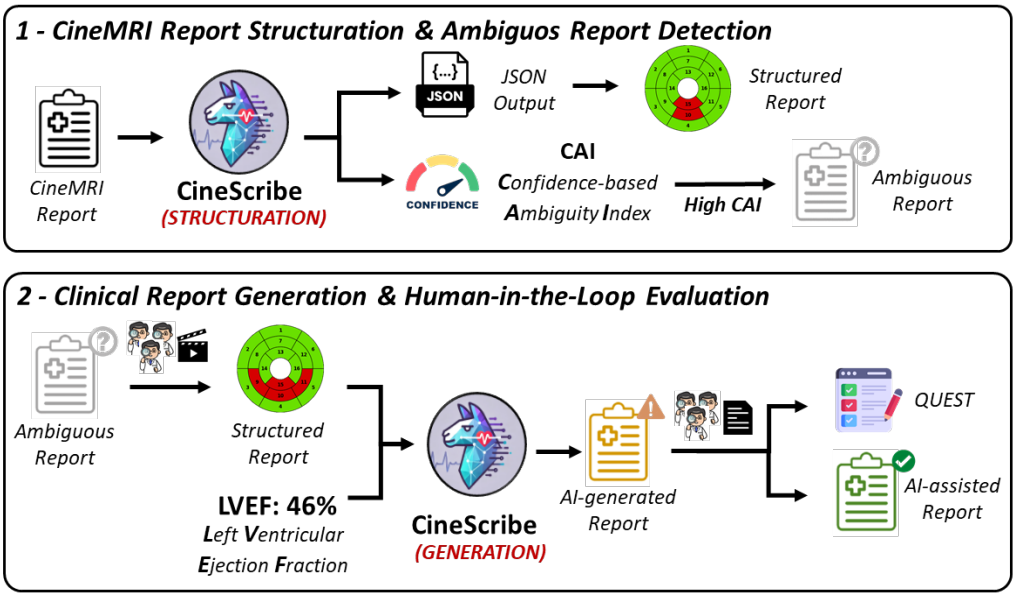

Figure 1. AI-assisted CineMRI Reporting Framework Overview: In the first stage, CineScribe structures clinical reports, flagging potentially ambiguous cases. In the second stage, experts reassess the corresponding cineMRI scans, providing the final structured diagnosis, which is then used by CineScribe to generate a standardized clinical report. Generated reports undergo expert review and are used to produce the final AI-assisted report.

 ISBN: 978-1-68558-362-0