CineScribe: LLM-based Detection and Mitigation of Ambiguity in Cine Cardiac Magnetic Resonance Reports

Guillermo Villanueva Benito[1], Martin L. Descalzo[2], Sandra Pujadas[2], Juan Fernandez[2], Matias Calandrelli[2] and Paula Petrone[3]

[1]Barcelone Institute for Global Health (ISGlobal), [2]Hospital de la Santa Creu i Sant Pau,
[3]Barcelona Supercomputing Center (BSC-CNS)

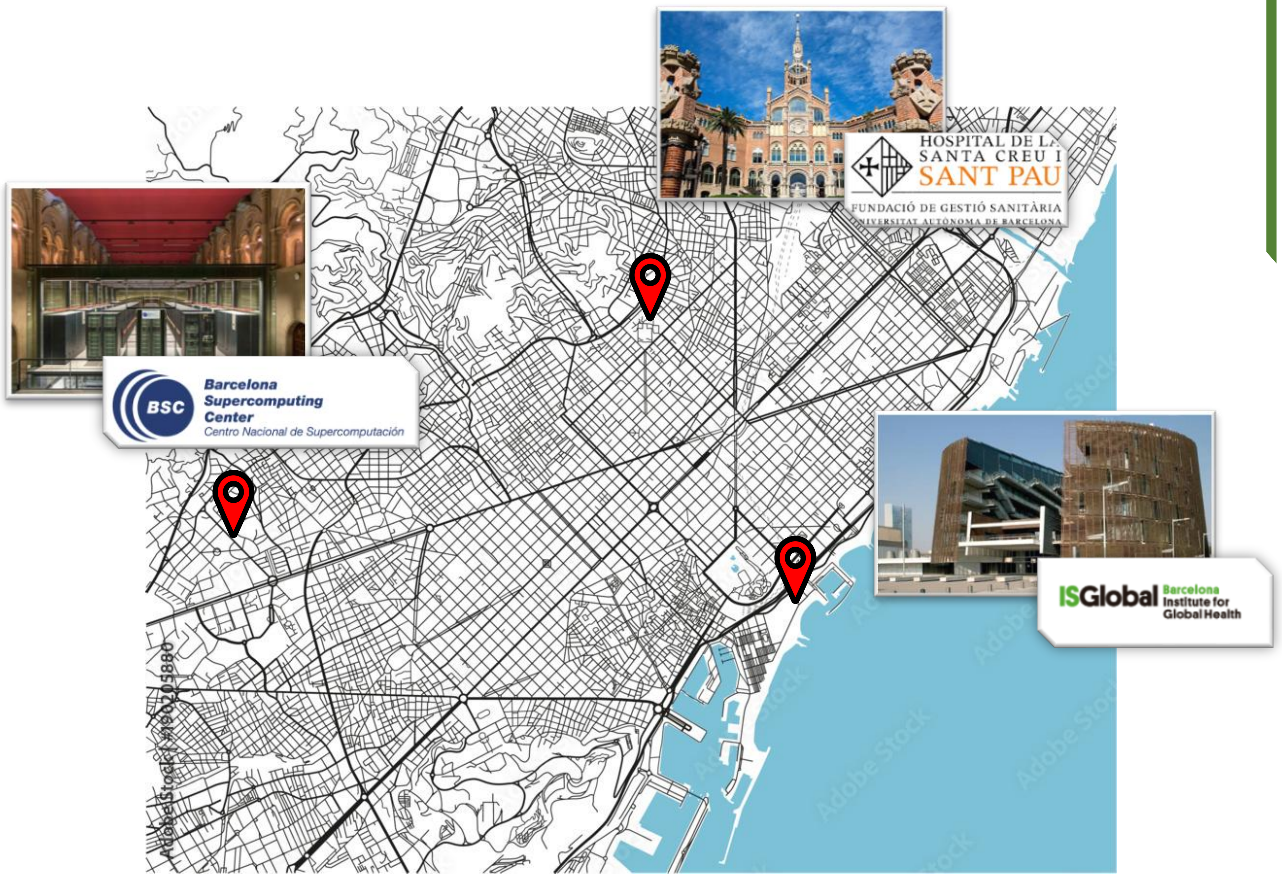
Contact email: **guillermo.villanueva@isglobal.org**

Guillermo Villanueva Benito

PhD Fellow | MSc in Applied Mathematics | BSc in Mathematics | BEng in Engineering Physics

Guillermo Villanueva is a PhD fellow in the **Digital Health Unit** at the Barcelona Institute for Global Health (ISGlobal) and Barcelona Supercomputing Center (BSC-CNS). He holds degrees in Mathematics and Engineering Physics from the Interdisciplinary Higher Education Center (CFIS) at the Universitat Politècnica de Catalunya (UPC).

His research interests lie in the intersection of artificial intelligence and healthcare, with a particular focus in real-word clinical impact.


HOSPITAL DE LA SANTA CREU I SANT PAU
FUNDACIÓ DE GESTIÓ SANITÀRIA
UNIVERSITAT AUTÒNOMA DE BARCELONA
BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
ISGlobal Barcelona Institute for Global Health

BARCELONA

1. Aims and Contributions of Our Paper

Towards AI-enhanced cineCMR documentation

In our paper, we aimed at:

Develop and evaluate a LLM-based framework in cineCMR reporting that:

1. Extracts structured diagnostic information from clinical report.
2. Identifies ambiguous clinical reports using model-derived confidence scores.
3. Supports standardized cineCMR documentation from expert-validated diagnostic findings.

This work makes four key contributions:

1. Ambiguity in cineCMR reports exists and is primarily driven by diagnostic complexity.
2. CineScribe enables accurate extraction of left ventricular Regional Wall Motion Abnormalities (RWMA) from free-text reports.
3. Reduced model confidence during structured information extraction serves as a novel predictor of clinical text ambiguity.
4. A human-in-the-loop framework that prioritizes complex cases for expert consensus review and supports standardized final report generation.

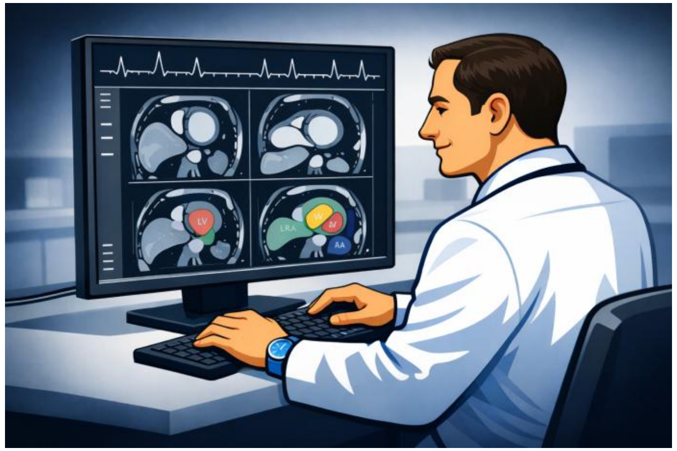
2. Problem and Unmet Need

Subjective and time-consuming diagnoses lead to severe patient outcomes

Cardiovascular diseases (CVDs) are the leading cause of death globally (31% of global deaths)

Left Ventricular Regional Wall Motion Abnormalities (RWMAs) describe kinetic alterations of the left ventricular heart muscle

Cine Cardiac Magnetic Resonance (cineCMR)



Visual Assessment and Diagnosis

- **Complex and subjective diagnoses** from human interpretation
- **Inconsistent diagnosis** due to inter-observer variability

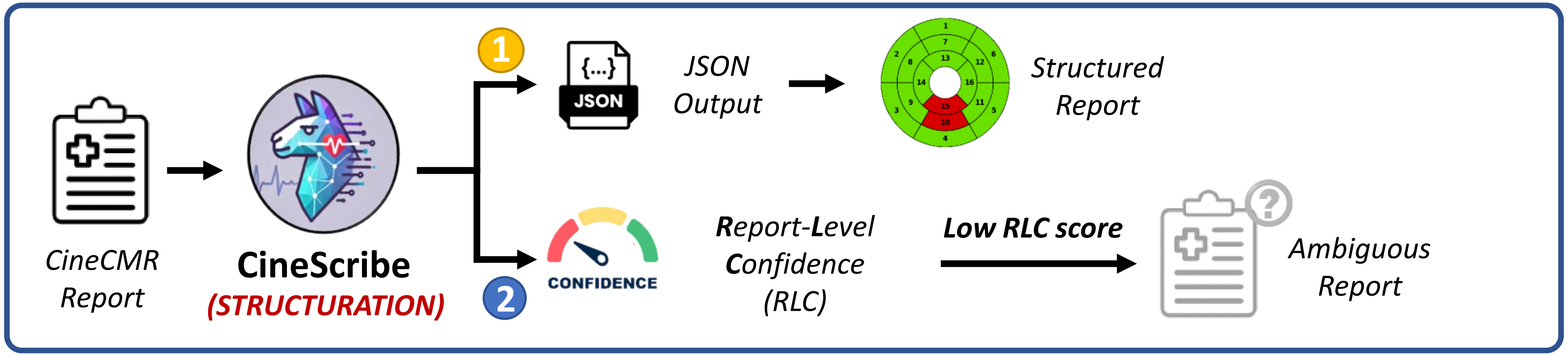
Manual Entry Documentation

- **High workload**
- **Low standardization**
- Prone to **ambiguities** and miscommunication

3.2. CineScribe: A domain-adapted lightweight LLM

Built on Llama3-8B* using Low-Rank Adaptation (LoRA)

CineCMR Report Structuration
&
Confidence-Based Ambiguity Detection

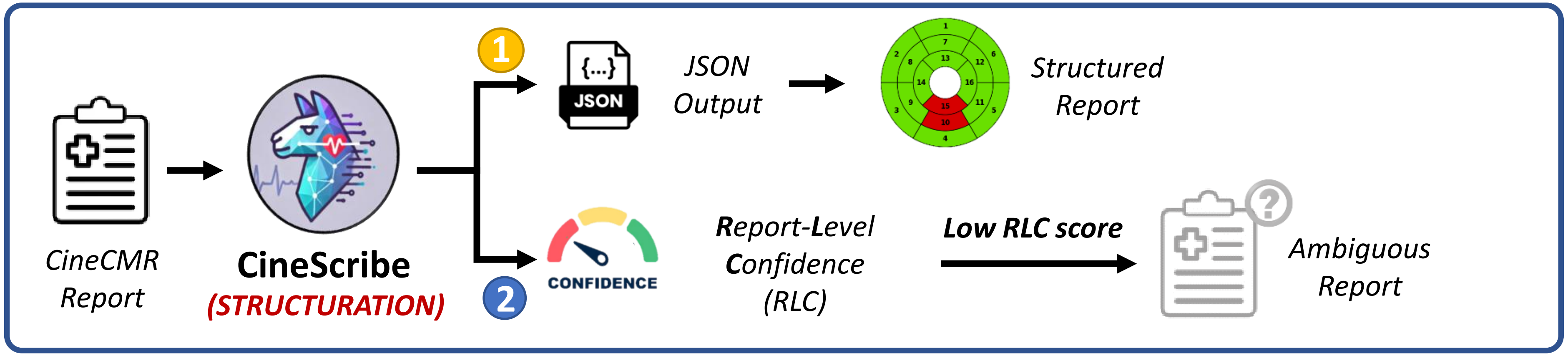


*AI@Meta

3.2. CineScribe: A domain-adapted lightweight LLM

Built on Llama3-8B* using Low-Rank Adaptation (LoRA)

CineCMR Report Structuration
&
Confidence-Based Ambiguity Detection



CineCMR Report

CineScribe
(STRUCTURATION)

1

{...}
JSON

JSON Output

Structured Report

2

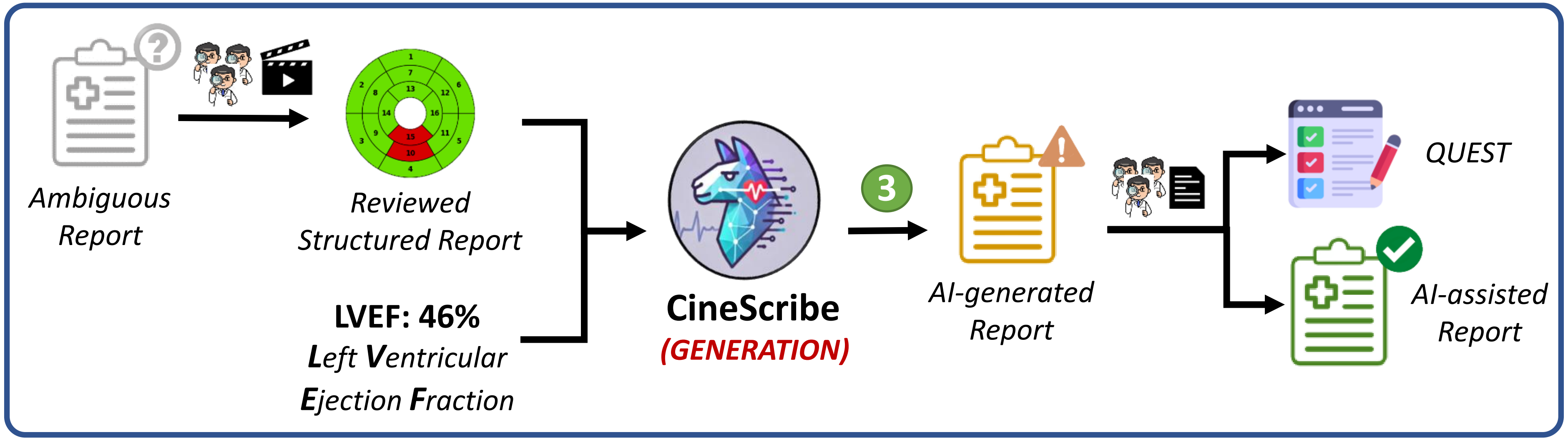
CONFIDENCE

Report-Level Confidence (RLC)

Low RLC score

Ambiguous Report

AI-assisted Report Generation and Evaluation



Ambiguous Report

Reviewed Structured Report

LVEF: 46%
***L**eft **V**entricular*
***E**jection **F**raction*

CineScribe
(GENERATION)

3

AI-generated Report

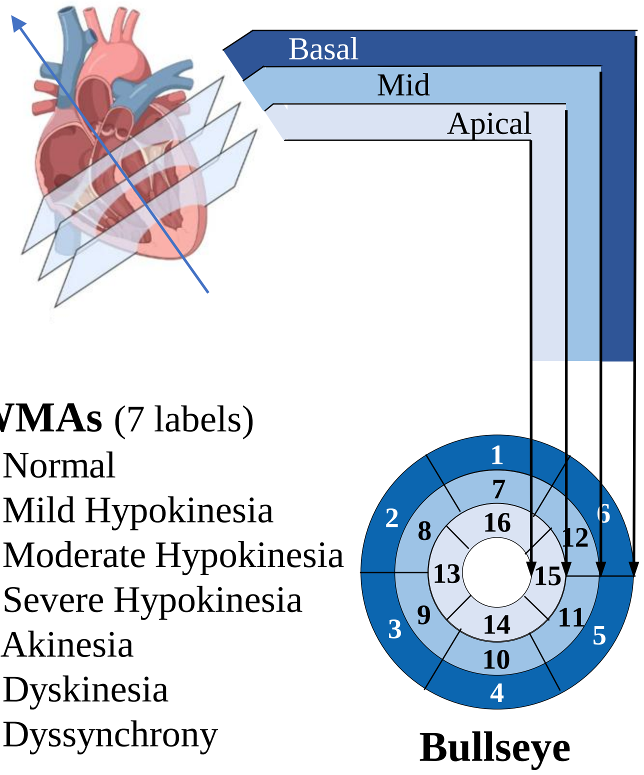
QUEST

AI-assisted Report

*AI@Meta

3.1. CineCMR Report Structuration (RWMAs)

Ambiguous reports lead to inter-annotator disagreement during report structuration



RWMAs (7 labels)

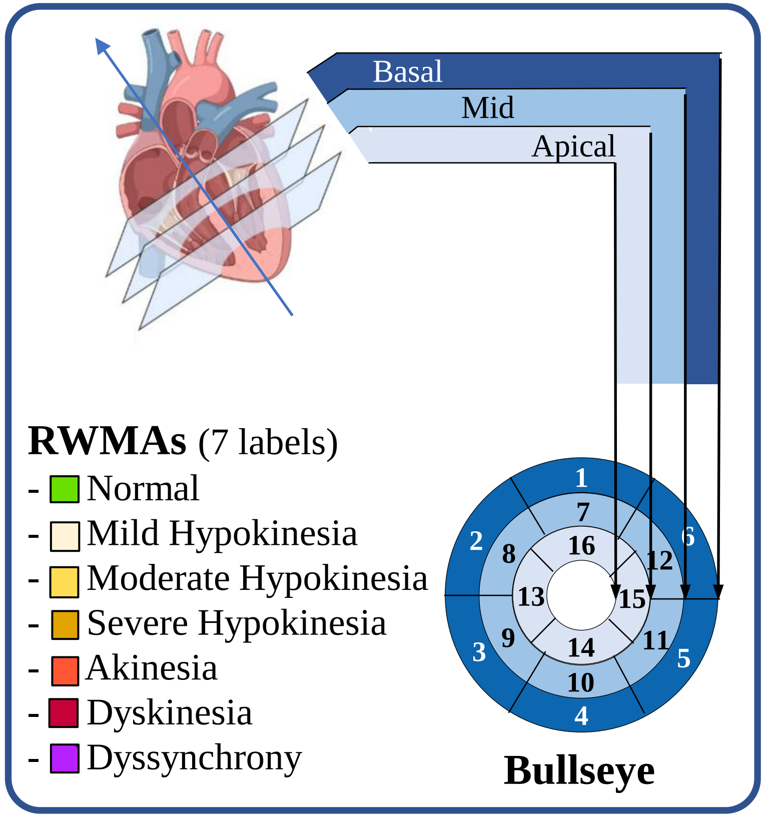
- 🟩 Normal
- ⬜ Mild Hypokinesia
- 🟨 Moderate Hypokinesia
- 🟧 Severe Hypokinesia
- 🟥 Akinesia
- 🟥 Dyskinesia
- 🟪 Dyssynchrony

16-segment AHA model*

*Cerqueira, M. D., et al

3.1. CineCMR Report Structuration (RWMAs)

Ambiguous reports lead to inter-annotator disagreement during report structuration



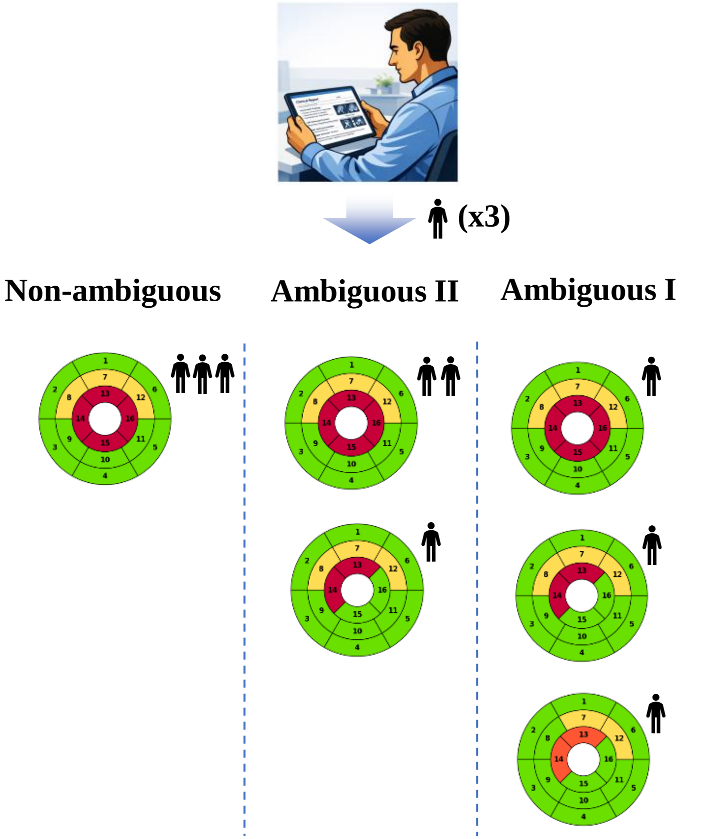
RWMAs (7 labels)

- 🟩 Normal
- □ Mild Hypokinesia
- 🟨 Moderate Hypokinesia
- 🟧 Severe Hypokinesia
- 🟥 Akinesia
- 🟥 Dyskinesia
- 🟪 Dyssynchrony

16-segment AHA model*

1. Report Structuration

(Interpretation from the text)

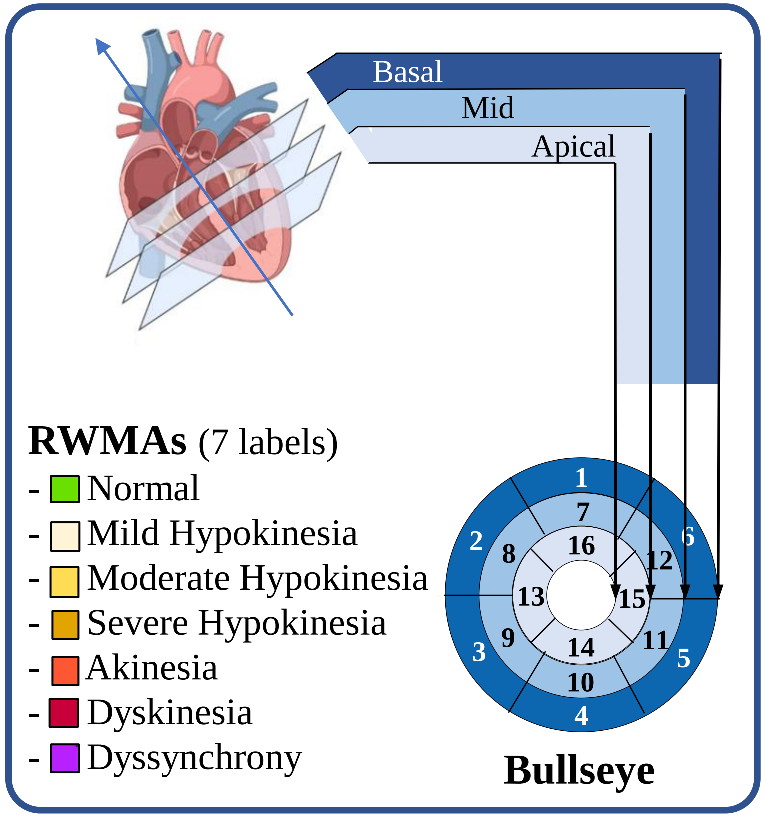


Higher Inter-Annotator Disagreement

*Cerqueira, M. D., et al

3.1. CineCMR Report Structuration (RWMAs)

Ambiguous reports lead to inter-annotator disagreement during report structuration



RWMAs (7 labels)

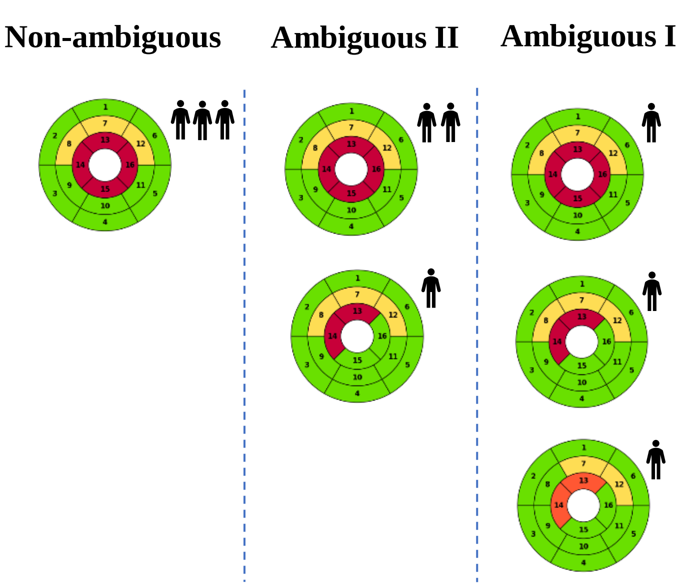
- 🟩 Normal
- ⬜ Mild Hypokinesia
- 🟨 Moderate Hypokinesia
- 🟧 Severe Hypokinesia
- 🟥 Akinesia
- 🟥 Dyskinesia
- 🟪 Dyssynchrony

16-segment AHA model*

1. Report Structuration

(Interpretation from the text)

(x3)

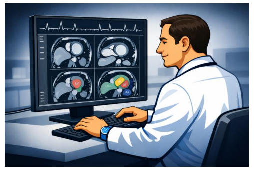


Non-ambiguous **Ambiguous II** **Ambiguous I**

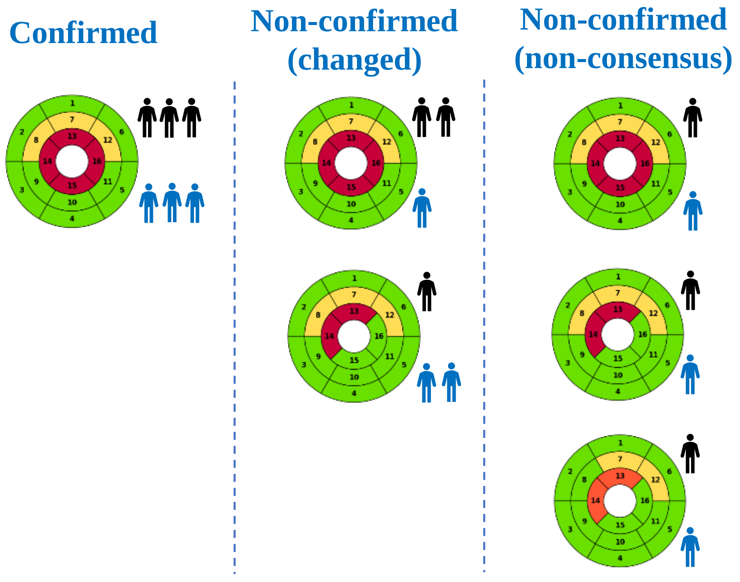
Higher Inter-Annotator Disagreement

2. CineCMR Review

(Verification based on the data)



(x3)



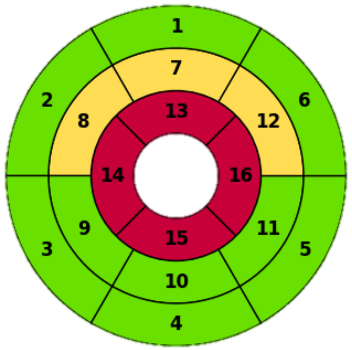
Confirmed **Non-confirmed (changed)** **Non-confirmed (non-consensus)**

Higher Diagnostic Complexity

*Cerqueira, M. D., et al

3.3. Report-Level Confidence (RLC)

A global confidence score for structured (JSON) generated outputs



Bullseye

Predicted Labels (j=1,…,16)

```
{"Anterior_basal":        "Normal",
"Anteroseptal_basal":     "Normal",
"Inferoseptal_basal":     "Normal",
"Inferior_basal":         "Normal",
 "Inferolateral_basal":   "Normal",
"Anterolateral_basal":    "Normal",
"Anterior_medio":         "Hipocinesia moderada",
"Anteroseptal_medio":     "Hipocinesia moderada",
"Inferoseptal_medio":     "Normal",
"Inferior_medio":         "Normal",
"Inferolateral_medio":    "Normal",
"Anterolateral_medio":    "Hipocinesia moderada",
"Anterior_apical":        "Discinesia",
"Septal_apical":          "Discinesia",
"Inferior_apical":        "Discinesia",
"Lateral_apical":         "Discinesia"}
```

Structured Report (JSON)

Predicted Label j

$$\text{token}_1|\text{token}_2|\dots|\text{token}_{N_j}$$

Token-level log-probabilities

$$\log p_1|\log p_2|\dots|\log p_{N_j}$$

Minimum

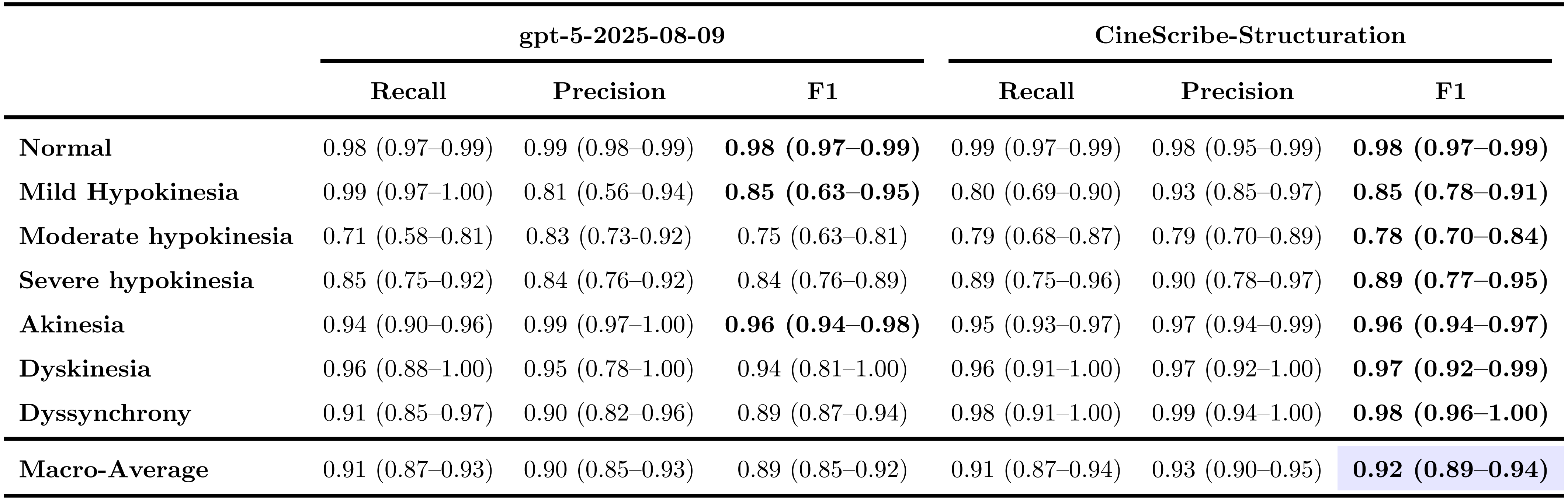
$$\log P_{\text{Label}_j} = \min_{i\in\{1,\dots,N_j\}} \log p_i$$

$$RLC = e^{\sum_{j=1}^{16} \log P_{\text{Label}_j}} = \prod_{j=1}^{16} P_{\text{Label}_j}$$

Report-Level Confidence (RLC)

4.1. Report Structuration Performance

High report structuration performance across RWMAs labels



	gpt-5-2025-08-09			CineScribe-Structuration		
	Recall	Precision	F1	Recall	Precision	F1
Normal	0.98 (0.97–0.99)	0.99 (0.98–0.99)	**0.98 (0.97–0.99)**	0.99 (0.97–0.99)	0.98 (0.95–0.99)	**0.98 (0.97–0.99)**
Mild Hypokinesia	0.99 (0.97–1.00)	0.81 (0.56–0.94)	**0.85 (0.63–0.95)**	0.80 (0.69–0.90)	0.93 (0.85–0.97)	**0.85 (0.78–0.91)**
Moderate hypokinesia	0.71 (0.58–0.81)	0.83 (0.73-0.92)	0.75 (0.63–0.81)	0.79 (0.68–0.87)	0.79 (0.70–0.89)	**0.78 (0.70–0.84)**
Severe hypokinesia	0.85 (0.75–0.92)	0.84 (0.76–0.92)	0.84 (0.76–0.89)	0.89 (0.75–0.96)	0.90 (0.78–0.97)	**0.89 (0.77–0.95)**
Akinesia	0.94 (0.90–0.96)	0.99 (0.97–1.00)	**0.96 (0.94–0.98)**	0.95 (0.93–0.97)	0.97 (0.94–0.99)	**0.96 (0.94–0.97)**
Dyskinesia	0.96 (0.88–1.00)	0.95 (0.78–1.00)	0.94 (0.81–1.00)	0.96 (0.91–1.00)	0.97 (0.92–1.00)	**0.97 (0.92–0.99)**
Dyssynchrony	0.91 (0.85–0.97)	0.90 (0.82–0.96)	0.89 (0.87–0.94)	0.98 (0.91–1.00)	0.99 (0.94–1.00)	**0.98 (0.96–1.00)**
Macro-Average	0.91 (0.87–0.93)	0.90 (0.85–0.93)	0.89 (0.85–0.92)	0.91 (0.87–0.94)	0.93 (0.90–0.95)	**0.92 (0.89–0.94)**

Table 1. Report-Structuration performance comparison with GPT-5 (zero-shot).

4.2. Confidence-Based Ambiguity Detection

Report-Level Confidence (RLC) predicts ambiguous reports and reveals underlying diagnostic complexity

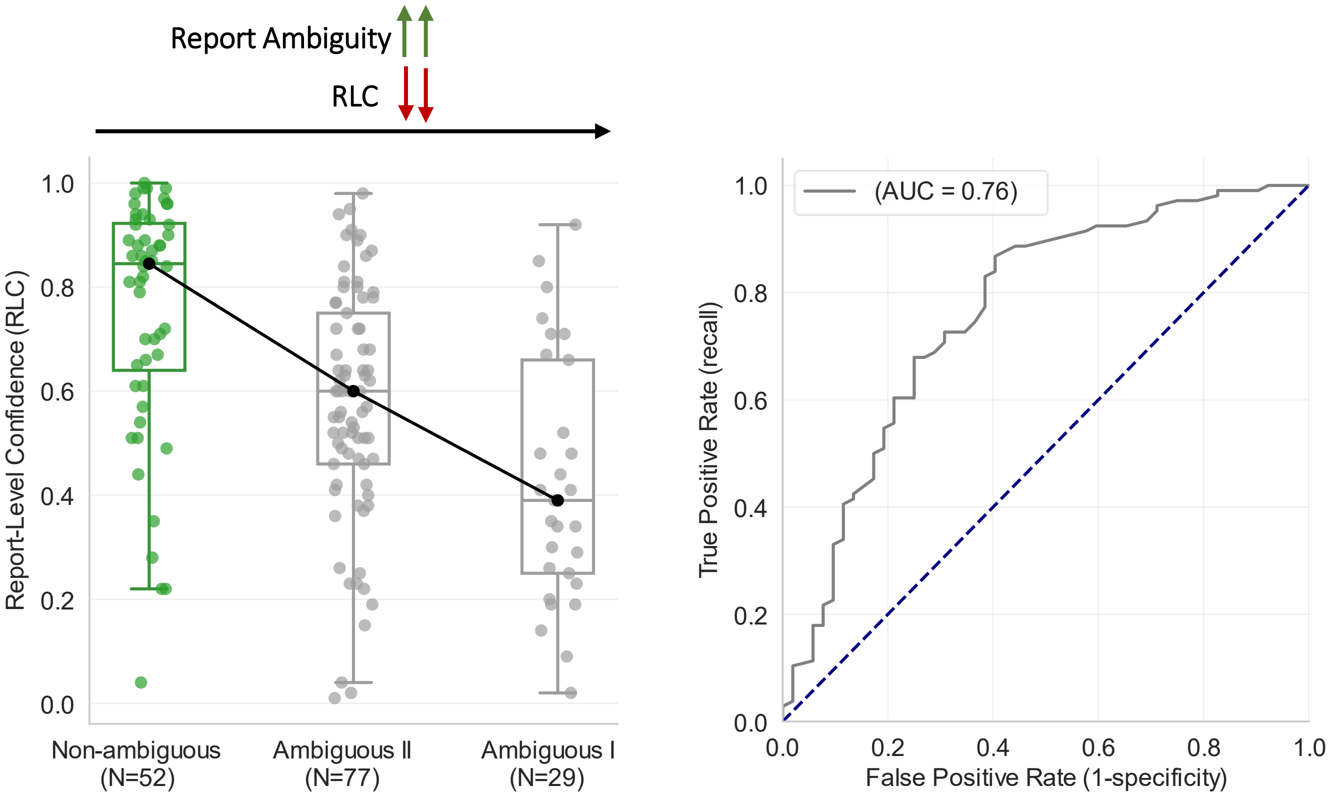


Figure 1. Ambiguous report detection

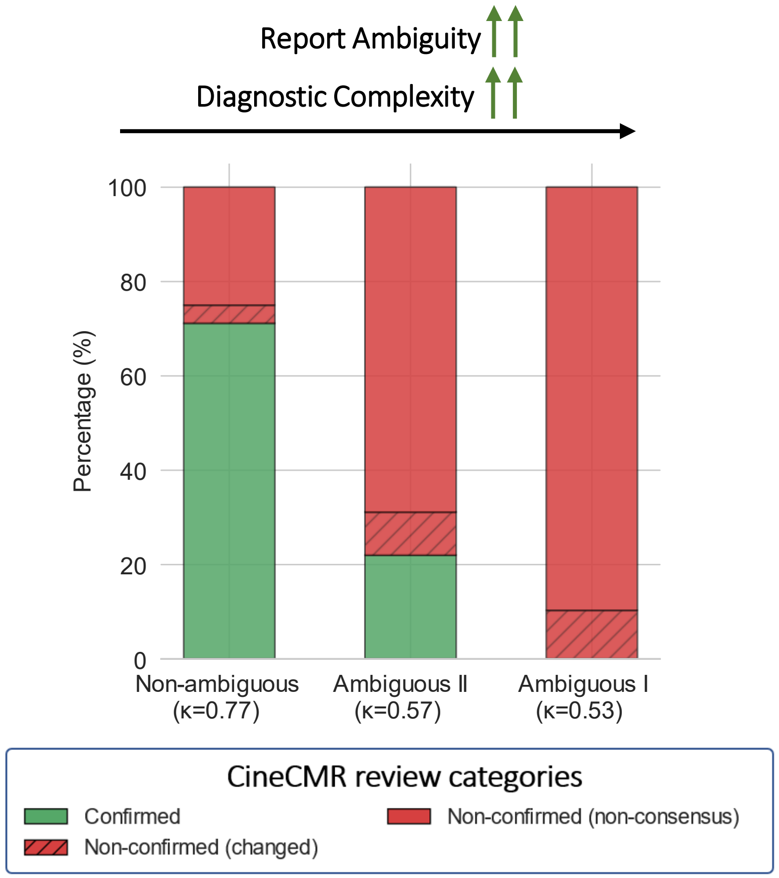


Figure 2. CineCMR review

4.3. Expert Evaluation of Automated Report Generation

Remarkable clinical report generation capabilities supporting cineCMR documentation

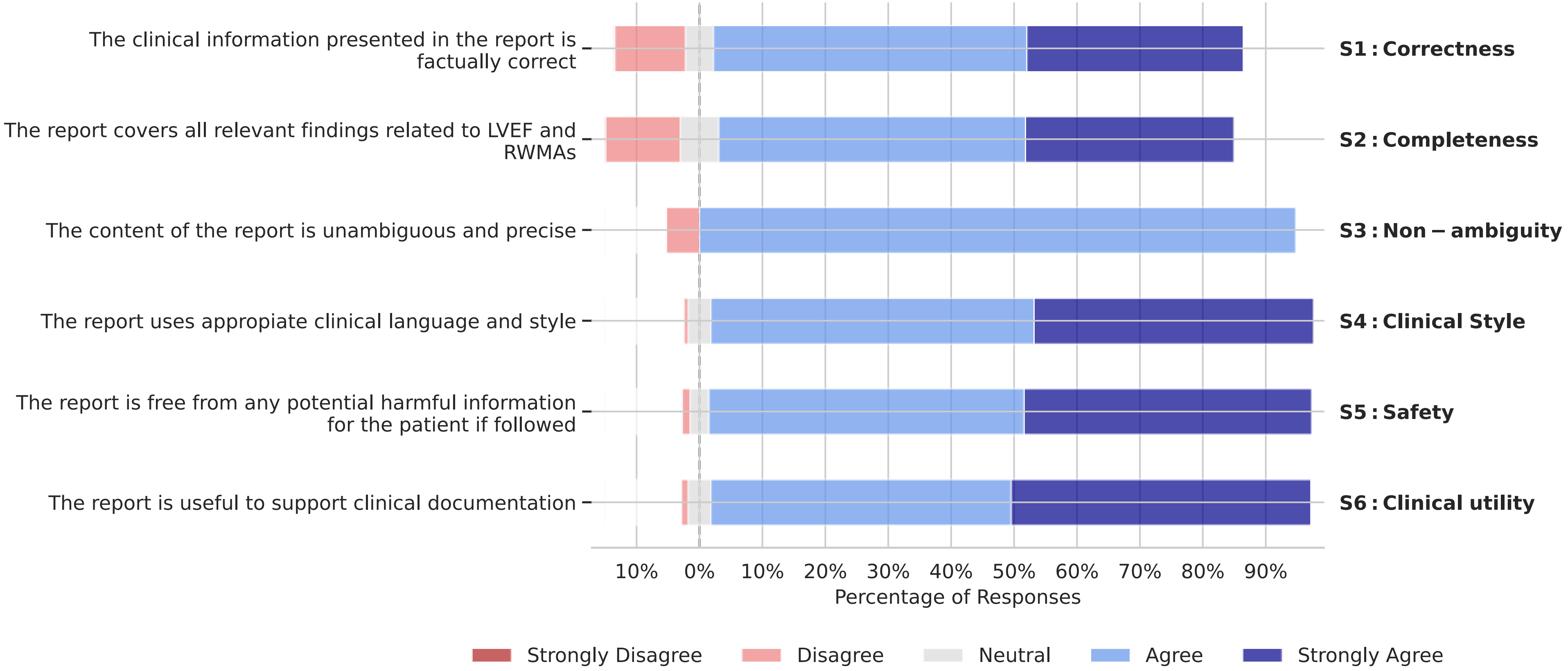
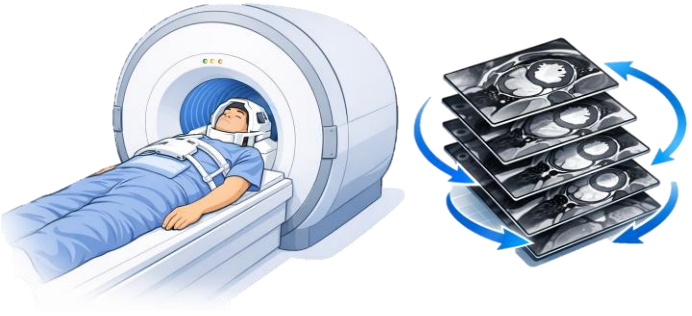


Figure 3. Clinical Evaluation of CineScribe generated reports following QUEST (5-point Likert)

4.4. A Human-in-the-Loop Approach

From prioritized cineCMR review to standardized AI-assisted report generation



Complex CineCMR Diagnostic Case
(Prone to inter-observer variability and report ambiguity)

Structured-Guided
Expert Consensus Review

AI-assisted
Entry Documentation

5. Conclusion and Future Work

Towards AI-enhanced cineCMR documentation

Conclusion:

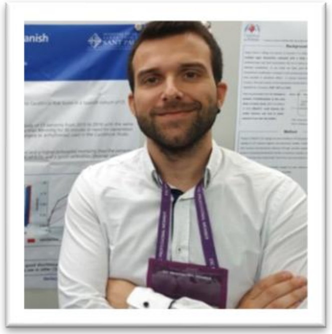
1. Our work introduces LLM-derived model's confidence during clinical information extraction as a novel predictor for report text ambiguity.
2. CineScribe provides a practical approach to identify complex diagnostic cases that may benefit from expert consensus review, thereby promoting more reliable diagnosis
3. CineScribe report generation capabilities illustrates a path towards more standardized and automated cineMRI documentation.

Future work:

1. Further explore human-in-the-loop approaches, such as the one proposed in this work, to improve report quality and support continued fine-tuning and enhancement of the model's capabilities.
2. Further directions also include the integration of image-derived cineCMR features, as well as prospective deployment and validation in multi-center settings to assess real-time performance, robustness, and clinical impact.
3. Finally, continued research is needed to better characterize and address the intrinsic sources of variability in cineCMR assessment, particularly in diagnostically complex cases.

Acknowledgments

Co-designing AI-Enhanced cineCMR Diagnosis in Cardiology



BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

CineScribe: LLM-based Detection and Mitigation of Ambiguity in Cine Cardiac Magnetic Resonance Reports

Guillermo Villanueva Benito[1], Martin L. Descalzo[2], Sandra Pujadas[2], Juan Fernandez[2], Matias Calandrelli[2] and Paula Petrone[3]

[1]Barcelone Institute for Global Health (ISGlobal), [2]Hospital de la Santa Creu i Sant Pau,
[3]Barcelona Supercomputing Center (BSC-CNS)

Contact email: **guillermo.villanueva@isglobal.org**