ACL 2026 SAN DIEGO
JULY 2–7

Learning from Textual Radiology Reports: A Benchmark Dataset for Coronary CT Angiography

Sudharshan Balaji[1], Zhiyu Liu[5], Zhengyuan Jiang[1], Shuo Lei[2], Yimin Chen[3], Yang Xiao[4], Shone O. Almeida[1], Mathew Joseph Karivelil[1], Christopher Malanga[1], Ning Wang[1]

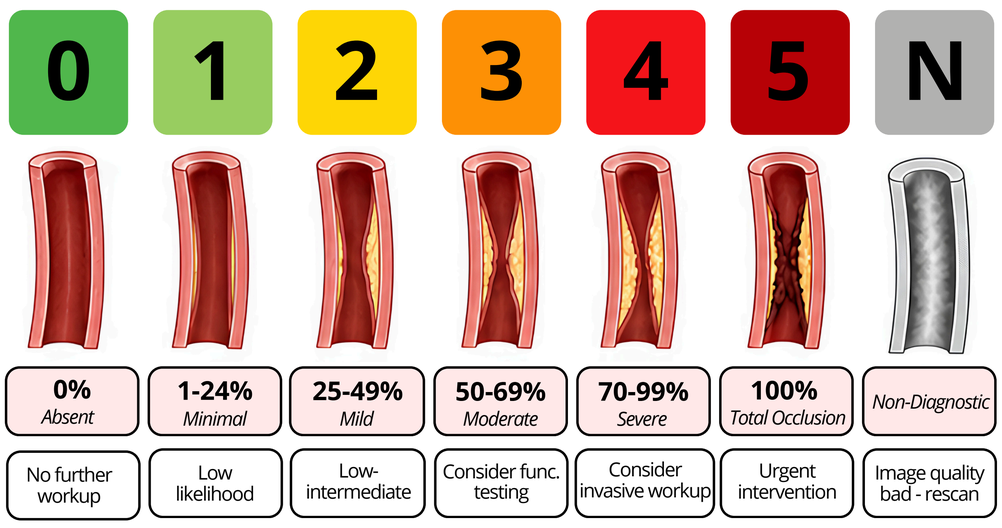
[1]University of South Florida, [2]Virginia Tech, [3]University of Massachusetts at Lowell, [4]University of Kentucky, [5]Baystate Medical Center

University of South Florida · Virginia Tech · University of Massachusetts Lowell · University of Kentucky · Baystate Health

1 Introduction to CADRADS

CCTA is the primary non-invasive scan for Coronary Artery Disease (CAD). After every scan, radiologists assign a CAD-RADS score (0-5 or N) that directly guides patient management

CAD-RADS Score Guide



Why automation matters here?

- Scores vary between readers and create bottlenecks at scale
- The output is a structured label that should be automatable
- Reliable automation standardizes scoring, reduces variability, and frees clinician time

2 Images vs Text Reports

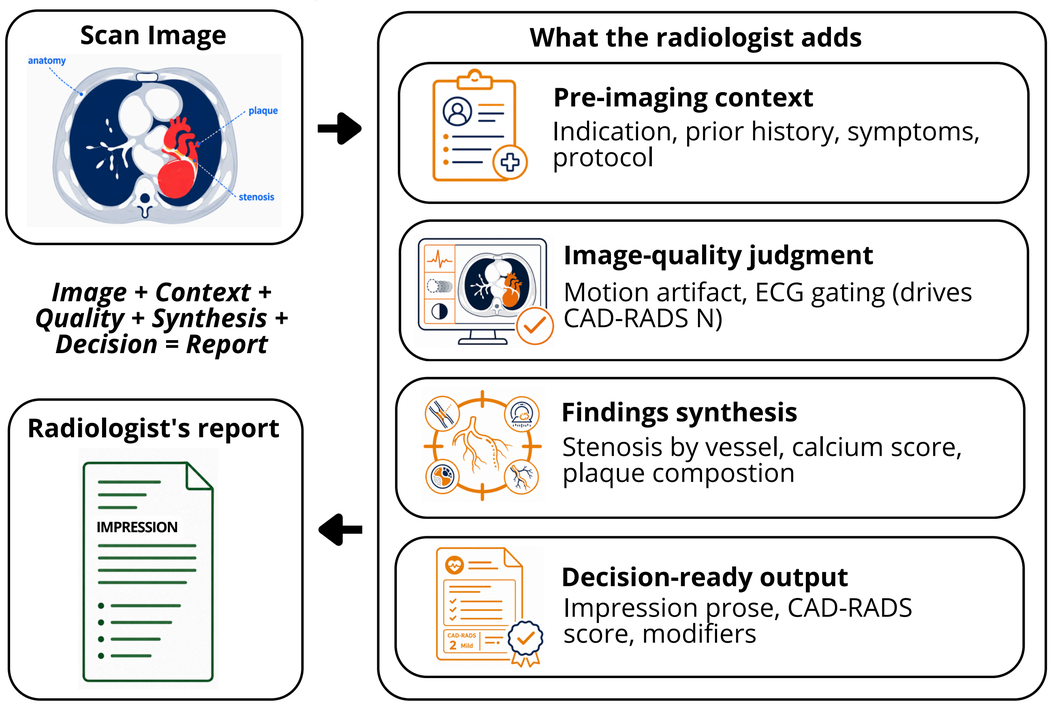
Image datasets for CAD-RADS exist and are valuable (ImageCAS, CorArtTS, CCA-200). They serve one point in the clinical workflow. Reports serve a different, broader one

Who reads it?	When useful	Cost / access
Radiologists	During scan interpretation	GB-scale PACS data
Clinicians, ED, payers	After the report enters EHR	Already indexed text

Image-based methods support the radiologist at read time. Report-based methods standardize what reaches everyone else

What is in the report?

A CCTA report has four functional layers. Only the first is image-derivable; the other three are radiologist-added information that image-based methods cannot recover



The CAD-RADS score is the result of all four layers. The report is the only artifact where the **full reasoning is captured in one place**

3 Same diagnosis, different words

Two reports with identical CAD-RADS scores can share almost no surface form

Report 5 (CAD-RADS 1)

LCx - Findings: The **LCx** was poorly.....segment resulting in **less than 25%** luminal stenosis.

LAD - Findings: The **left anterior descending** artery was poorly opacified....

Impression: There is ... in the proximal **left circumflex coronary artery** resulting in **less than 25%** stenosis

Report 9 (CAD-RADS 1)

LCx - Findings: The **left circumflex** artery is patent with **<24%** calcified plaque proximally

LAD - Findings: Less than **24%** low ... with flecks of calcification in proximal LAD.

Impression: **Less than 24%** ... of calcification in proximal LAD. **Less than 24%** low attenuation plaques in ... **RCA.**

What varies within one instution

- **Vocabulary:** "LCx" vs "left circumflex artery"
- **Vessel coverage:** LCx + LAD vs LCx + LAD + RCA
- **Stenosis encoding:** "less than 25%" vs " <24%"
- **Section structure:** Same findings, different locations

4 Meet CCTA-RADS

Largest publicly available, real-world dataset for CAD-RADS scoring from clinical text

- 940 reports from Tampa General Hospital
- 2020–2024; CAD-RADS 1.0 and 2.0
- Multiple authors - radiologists/cardiologists
- IRB-approved, HIPAA de-identified

Reports **940** · Vocabulary **6,798** words · Avg/median len. **529 / 518** tokens · Avg sentences **42**

Min / Max length: **309** words – **1020** words

Label distribution (CAD-RADS class)

0	1	2	3	4	5	N
338 (36.0%)	140 (14.9%)	134 (14.3%)	74 (7.9%)	72 (7.7%)	7 (0.7%)	175 (18.6%)

Study	Size	Type	Dataset	Code	Weights	Approach
Wang et al. (2024)	590	Real world	✗	✗	✗	Direct LLM (GPT-3.5/4, Llama3)
Arnold et al. (2025)	200	Synthetic	✗	✗	✗	Direct LLM (GPT-4o, Llama3-70b)
Min et al. (2025)	319	Real world	✓	✗	✗	Direct LLM (GPT-4)
Ours	**940**	**Real world**	✓	✓	✓	**Two-stage**

5 Direct Models Struggle with Clinical Heterogeneity

LLMs look competitive, but miss non-diagnostic cases, BERTs fail on raw heterogeneous text

Prompting	Model	Overall F1	CADRADS N Acc.
Zero-Shot	GPT-4o	82.76%	42.40%
	GPT-o3	83.49%	50.65%
One-Shot	GPT-4o	81.32%	42.17%
	GPT-o3	83.90%	49.57%
CoT	GPT-4o	81.85%	46.75%
	GPT-o3	84.08%	53.91%

Model	Overall F1	CADRADS 4 Acc.	CADRADS 5 Acc.
BERT	73.00%	0.00%	0.00%
BioBERT	70.77%	41.14%	0.00%
ClinicalBERT	76.74%	66.67%	0.00%
Bio+Clnc.BERT	77.31%	88.89%	0.00%

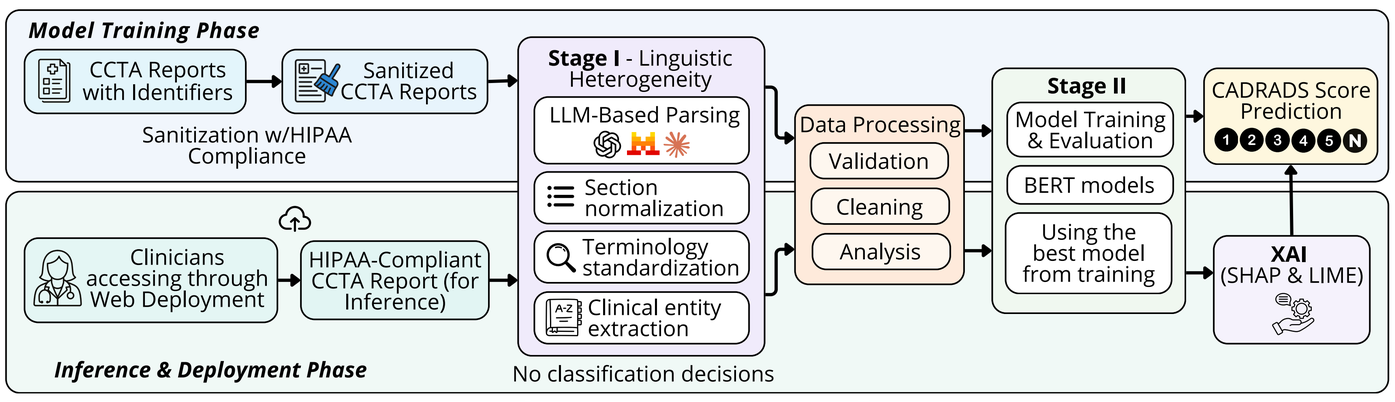
Overall F1 looks acceptable, but class-wise performance breaks

N Performance CAD-RADS N remains weak, best direct LLM accuracy = 53.91%

Structural noise Clinical knowledge is not enough when the input structure is noisy

The bottleneck is **task design**: one model is forced to **parse** and **reason** at once

6 Separating the jobs - Two stage pipeline



Head-to-head

Approach	Best model	F1	Δ vs Two-stage
Direct LLM	GPT-o3 CoT	84.08%	-6.04
Raw-text BERT	Bio+Clinical BERT	77.31%	-12.81
Two-stage (ours)	**BioBERT**	**90.12%**	-

Full two-stage results

Classifier	F1	Δ vs Raw
BERT	84.08	+11.08
ClinicalBERT	86.82	+10.08
BioBERT	**90.12**	**+19.35**
Bio+Clinical BERT	89.31	+12.00

Class-wise - where it matters most clinically

CAD-RADS	BioBERT (two-stage)	GPT-o3 CoT (direct)	Δ
0	100.00%	96.78%	+3.2
1	95.65%	94.70%	+0.95
2	91.67%	93.75%	-2.1
3	**100.00%**	**91.86%**	**+8.1**
4	**100.00%**	**92.22%**	**+7.8**
5	0.00% (n=7)	82.35%	-
N	**68.18%**	**53.91%**	**+14.3**

Structured outputs are consistent: BERTScore F1 97.47% across reports with vs. without explicit scores; cross-LLM agreement >0.93

Decoupling works Every BERT improves after structuring, best F1 = 90.12%

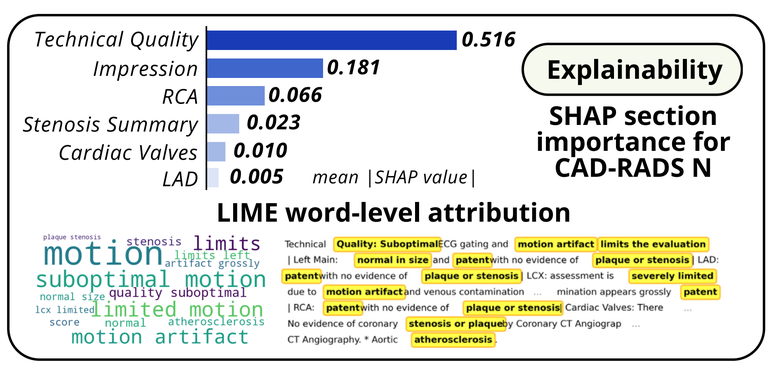
Clinical gains CAD-RADS 3/4 = 100%; CAD-RADS N improves +14.3.

Errors are bound Most mistakes are near-misses along the severity scale

Explanations align Technical Quality drives CAD-RADS N predictions (↓)

BioBERT confusion matrix (test set)

Tr↓\Pd→	0	1	2	3	4	5	N
0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
1	4.3	95.7	0.0	0.0	0.0	0.0	0.0
2	0.0	4.2	91.7	4.2	0.0	0.0	0.0
3	0.0	0.0	0.0	100.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	100.0	0.0	0.0
5	0.0	100.0	0.0	0.0	0.0	0.0	0.0
N	9.1	13.6	4.5	4.5	0.0	0.0	68.2



Three takeaways

1 Real-world clinical text is harder than benchmark text suggests

Direct LLMs and raw-text BERT models look competitive at a headline level, but break on heterogeneous CCTA reports

*The problem is not just medical knowledge, it is **surface variation, inconsistent structure**, and clinically important cues like technical quality*

2 Decoupling structuring from reasoning consistently wins

Raw Report Text → Stage 1 (Parser) → Stage 2 (Classify) → Better Classification

*Structuring step is **key**, the classifier choice **secondary***

3 CCTA-RADS enables reproducibility

Dataset · Code · Structured Representations · Trained Weights

Limitations we own

- Single-institution data from Tampa General
- CAD-RADS 5 remains rare and underpowered