Journal of Cardiovascular Computed Tomography xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Cardiovascular Computed Tomography

journal homepage: www.JournalofCardiovascularCT.com

Research Paper

Performance of large language models for CAD-RADS 2.0 classification derived from cardiac CT reports

Philipp Georg Arnold [a], Maximilian Frederik Russe [a], Fabian Bamberg [a], Tilman Emrich [b,c], Milán Vecsey-Nagy [c,d], Ayaat Ashi [f], Dmitrij Kravchenko [c,e], Ákos Varga-Szemes [c], Martin Soschynski [a], Alexander Rau [g], Elmar Kotter [a], Muhammad Taha Hagar [a,c,*]

[a] *Department of Diagnostic and Interventional Radiology, Medical Center, University of Freiburg, Faculty of Medicine, Freiburg, Germany*
[b] *Department of Diagnostic and Interventional Radiology, University Medical Center of the Johannes Gutenberg-University, Mainz, Germany*
[c] *Division of Cardiovascular Imaging, Department of Radiology and Radiological Science, Medical University of South Carolina, Charleston, USA*
[d] *Cardiovascular Imaging Research Group, Heart and Vascular Center, Semmelweis University, Budapest, Hungary*
[e] *Department of Diagnostic and Interventional Radiology, University Hospital Bonn, Bonn, Germany*
[f] *School of Medicine, Cardiff University, Neuadd Meirionnydd, Cardiff, United Kingdom*
[g] *Department of Neuroradiology, Medical Center, University of Freiburg, Faculty of Medicine, Freiburg, Germany*

ARTICLE INFO

Keywords:
Coronary artery disease
Cardiac CT
Artificial intelligence
Natural language processing
Clinical decision-making
Data security

ABSTRACT

Background: The Coronary Artery Disease-Reporting and Data System (CAD-RADS) 2.0 offers standardized guidelines for interpreting coronary artery disease in cardiac CT. Accurate and consistent CAD-RADS 2.0 scoring is crucial for comprehensive disease characterization and clinical decision-making. This study investigates the capability of large language models (LLMs) to autonomously generate CAD-RADS 2.0 scores from cardiac CT reports.

Methods: A dataset of cardiac CT reports was created to evaluate the performance of several state-of-the-art LLMs in generating CAD-RADS 2.0 scores via in-context learning. The tested models comprised GPT-3.5, GPT-4o, Mistral 7b, Mixtral 8 × 7b, Llama3 8b, Llama3 8b with a 64k context length, and Llama3 70b. The generated scores from each model were compared to the ground truth, which was provided by two board-certified cardiothoracic radiologists in consensus based on the reports.

Results: The final set comprised 200 cardiac CT reports. GPT-4o and Llama3 70b achieved the highest accuracy in generating full CAD-RADS 2.0 scores including all modifiers with a performance rate of 93 % and 92.5 %, respectively, followed by Mixtral 8 × 7b with 78 %. In contrast, older LLMs, such as Mistral 7b and GPT-3.5 performed poorly (16 %) and Llama3 8b demonstrated intermediate results with an accuracy of 41.5 %.

Conclusion: LLMs enhanced with in-context learning are capable of autonomously generating CAD-RADS 2.0 scores for cardiac CT reports with excellent accuracy, potentially enhancing both the efficiency and consistency of cardiac CT reporting. Open-source models not only deliver competitive accuracy but also present the benefit of local hosting, mitigating concerns around data security.

1. Introduction

Coronary artery disease (CAD) constitutes a leading cause of morbidity and mortality worldwide,[1] underscoring the importance of early diagnosis and precise characterization. Coronary CT angiography (CCTA) is a crucial non-invasive imaging modality for assessing CAD and is established as a first-line test in multiple international guidelines.[2,3] To ensure standardized reporting of CCTA findings, the Coronary Artery Disease-Reporting and Data System (CAD-RADS) was developed in 2016.[4] It provides a structured framework for grading coronary artery stenosis, thereby reducing interobserver variability, facilitating clinical decision-making and communication.

In recent years, the understanding of CAD characterization has shifted from merely measuring coronary stenosis to a broader focus on plaque burden, plaque characteristics and high-risk features associated with plaque rupture.[5–7] Therefore, the CAD-RADS 2.0 was introduced,[8] and additionally incorporates plaque burden quantification and additional patient management recommendations. For this, it includes metrics for

[*] Corresponding author. Medical University of South Carolina, Department of Radiology & Radiological Sciences, 25Courtenay Drive, Charleston, SC 29425, USA.
E-mail address: hagarmu@musc.edus (M.T. Hagar).

https://doi.org/10.1016/j.jcct.2025.03.007
Received 17 November 2024; Received in revised form 1 February 2025; Accepted 28 March 2025

Please cite this article as: Arnold PG et al., Performance of large language models for CAD-RADS 2.0 classification derived from cardiac CT reports, Journal of Cardiovascular Computed Tomography, https://doi.org/10.1016/j.jcct.2025.03.007

overall plaque burden, and assessments of ischemia through techniques such as fractional flow reserve derived from CT (FFR-CT) and CT myocardial perfusion imaging. These additions allow for a more nuanced evaluation of coronary lesions, improving the selection of patients for preventive therapies, invasive procedures, and long-term management strategies.[9] However, due to these advancements, the complexity of CAD-RADS 2.0 reports has increased hampering consistent and accurate scoring.[10] While CAD-RADS 2.0 scoring is straightforward for experienced readers, its multilayered structure and additional modifiers may introduce variability in complex cases. Ensuring consistency is essential, as errors in scoring may lead to misjudgment of disease severity, potentially impacting treatment decisions.[10] With increasing volumes of cardiac CT examinations,[11] there is a growing need for tools that assist radiologists in maintaining accuracy and consistency while reducing variability in CAD-RADS 2.0 scoring.

Recent advancements in artificial intelligence (AI)-based natural language processing might address this. Large language models (LLMs) such as GPT and Llama have demonstrated notable capabilities in understanding and generating complex medical narratives.[12–14] Recent studies provide evidence for the potential to summarize clinical reports and extract information from electronic health records.[15]

In contrast to these applications, the CAD-RADS 2.0 classification system is highly complex and multilayered. Our study evaluated the performance of state-of-the-art LLMs in generating CAD-RADS 2.0 scores from cardiac CT reports. Both proprietary systems, such as GPT-4o, and open-source models like Llama3, Mistral and Mixtral 8 × 7B were tested. The latter might allow for locally hosted solutions to mitigate privacy and data security concerns.

2. Methods

2.1. CCTA reports

To ensure data privacy and comply with cloud-based restrictions, a synthetic dataset of cardiac CT reports was generated. These reports were constructed based on CAD prevalence data from established literature,[16–18] while strictly adhering to the reporting template used at our institution, which aligns with international consensus guidelines (see supplements of the CAD-RADS 2.0 guideline).[8] Special care was taken to ensure they mirrored real-world clinical documentation, incorporating all relevant information for CAD-RADS classification while avoiding interpretative CT report elements to enable an objective assessment by LLMs.

To achieve this, cardiac CT reports were first extracted from retrospective cases, with all patient-identifiable information and impressions removed, leaving only objective findings. Then these reports were altered as follows: the synthetic reports were then semi-randomly generated by combining predefined text blocks for different sections, ensuring natural variability while preserving clinical consistency. If needed, stenosis severity, Agatston scores, and plaque burden values were assigned within clinically plausible ranges using a random generator. Each report underwent rigorous manual review and refinement by a board-certified radiologist (MTH) before the final CAD-RADS 2.0 classification was assigned in consensus by two board-certified radiologists (MTH, MS). This highly controlled approach was used to ensure that the dataset is not only synthetic but robust, standardized, and clinically representative. The methodology aligns with widely accepted practices in LLM research, where synthetic text data is routinely used to comply with privacy regulations while maintaining realistic evaluation environments for AI-driven applications in radiology.[12,13,19–21]

2.2. Sample size calculation

To ensure a robust statistical evaluation of the accuracy of LLMs in classifying CAD-RADS 2.0 categories, a sample size calculation was performed. The primary outcome of interest was the overall accuracy of LLMs compared to the ground truth derived from cardiac CT reports, with an accompanying 95 % confidence interval (CI). The required sample size (n) was calculated to ensure that the 95 % CI had a margin of error of 5 %. The estimated accuracy of the LLMs was assumed to be 90 %, based on contemporary literature evaluating the performance of state-of-the-art LLMs on structured classification tasks.[22] Using this value, the sample size was determined using the following equation for proportion estimation:

$$n = \frac{Z^2 \cdot (estimated\ accuracy) \cdot (1 - estimated\ accuracy)}{(margin\ of\ error)^2}$$

Substituting Z with 1.96 (for a 95 % CI range), the estimated accuracy at 90 %, and the margin of error at 5 %, a minimum of 138 reports were required. Using a more conservative assumption of 85 % accuracy, the necessary sample size equals 196. Therefore, 200 cardiac CT reports were used in this study.

2.3. CCTA interpretation according to CAD-RADS 2.0

As reference standard, the reports were classified according to the CAD-RADS 2.0 system by two board-certified radiologists (*MTH, MS*), with 6 and 9 years of experience in CCTA, respectively, in consensus. The format for CAD-RADS 2.0 coding followed the similar sequence as proposed in the guideline: *CAD-RADS/N/P/HRP/I/S/G/E*.[23] The LLMs` output was only considered correct if the complete CAD-RADS 2.0 code was provided in the correct sequence, including all applicable modifiers. This precision is critical, as further diagnostic testing, treatment recommendations, and adherence to structured reporting standards, are guided hereby.[23]

2.4. Prompt design, and task assignment

Precision prompts with in-context learning were employed to direct LLMs in application of the CAD-RADS 2.0 system on CT reports (Prompts provided in supplements). The provided context included information on the facts, that *CAD-RADS* scores range from 0 (no plaque or stenosis) to 5 (total occlusion), based on the diameter stenosis. CAD-RADS 1 represents minimal stenosis (1–24 %), CAD-RADS 2 indicates mild stenosis (25–49 %), and CAD-RADS 3 reflects moderate stenosis (50–69 %). CAD-RADS 4A represents severe stenosis (70–99 %) in one or two major vessels, while CAD-RADS 4B indicates left main stenosis of $\geq$50 % or severe three-vessel disease ($\geq$70 % stenosis in all three major coronary arteries).[23] The LLMs were also tasked with assessing overall coronary plaque burden using either the Agatston Score or Segment Involvement Score (SIS), as appropriate, and categorizing it into the distinct *P category*: P1 (mild, Agatston 1–100, SIS $\leq$2), P2 (moderate, 101–300, SIS 3–4), P3 (severe, 301–999, SIS 5–7), and P4 (extensive, $\geq$1000, SIS $\geq$8).[23] The models also evaluated the lacking interpretability of the entire study (CAD-RADS N) or part of a vessel (modifier N). In addition, the models evaluated High-Risk Plaque (modifier HRP) features within a plaque, requiring at least two of the following four components to be present: 1) positive remodeling, 2) low attenuation plaque (<30 Hounsfield Units), 3) spotty calcifications, and 4) the napkin-ring sign.[24,25] Furthermore, LLMs screened for functional assessments, such as FFR-CT, using specific thresholds: FFR-CT $\leq$ 0.75 indicated significant ischemia (modifier I+), values > 0.80 ruled out ischemia (I-), and values between 0.76 and 0.80 were classified as borderline or inconclusive (I±).[23] For CT perfusion, I+ modifier is assigned for reversible ischemia (perfusion defects during stress) and the I- modifier for the absence of ischemia or fixed myocardial infarcts, particularly for stenoses between 50 and 90 %, with further consideration for proximal lesions over 40 % or high-risk plaque features. The presence of stents or bypass grafts was flagged with S or G modifiers, and cases involving non-atherosclerotic causes of CAD, or anatomical variation of vessel origins, were marked as exceptions (E). GPT-3.5, GPT-4o, Llama 3 8B, Llama 3 8B with a 64k context length, Llama 3

70B, Mistral 7B, and Mixtral 8 × 7B were prompted similarly with identical prompts. An overview is shown in (Fig. 1), while characteristics of each used LLM is provided in Table 1.

2.5. Local LLM deployment

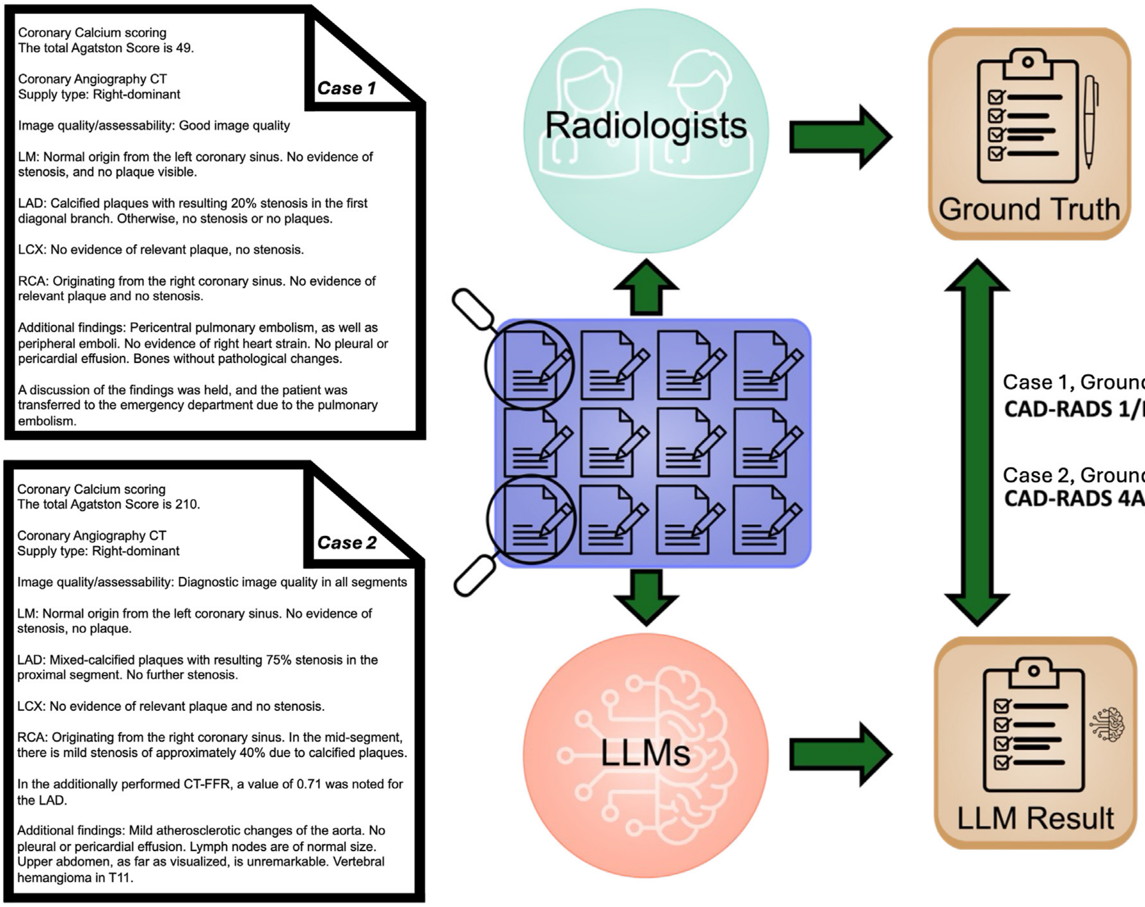
For local deployment of LLMs, we utilized LocalAI (https://localai.io), an open-source framework designed to run and serve models efficiently on local infrastructure. LocalAI was configured to function as an OpenAI-compatible API endpoint, allowing seamless integration with our experimental pipeline. The models were manually installed from repositories like Hugging Face (https://huggingface.co). The deployment was performed on a workstation equipped with an NVIDIA RTX 6000 GPU, which provided the necessary computational power and substantial VRAM required for efficient inference.

To optimize performance and reduce memory requirements, all models were deployed in a quantized GGUF (GPTQ for GGML Unified Format)version, which allows for reduced precision representations of model weights. Quantization compresses model parameters—e.g., from 16-bit floating point (FP16) to lower-bit integer formats (such as 4-bit, 5-bit, or 8-bit)—significantly reducing GPU memory requirements while maintaining inference performance.

We evaluated several locally hostable models with varying computational requirements. Llama 3 70B (∼48 GB GPU RAM, 4-bit quantization, 4096 context window), Llama 3 8B (∼6 GB GPU RAM, 5-bit quantization, 4096 context window), Llama 3 8B (64k context) (∼12 GB GPU RAM, 8-bit quantization, 65536 context window), Mistral 7B (∼6 GB GPU RAM, 4-bit quantization, 8192 context window), and Mixtral 8 × 7B (∼48 GB GPU RAM, 6-bit quantization, 32768 context window) were tested.

2.6. Statistics

Statistical analyses were performed using R (version 4.4.1). The assumption of normal distribution was assessed with the Shapiro-Wilk test. Quantitative variables were expressed as mean ± standard deviation (SD) for normally distributed data, and as median and interquartile range (IQR) for non-normal distributions. Categorical variables were presented as counts and percentages. We calculated Krippendorff's alpha to assess the inter-rater reliability between each LLM and the ground truth provided by two board-certified radiologists regarding the CAD-RADS score and Plaque assessment. Given the structured nature of CAD-RADS classification, it is essential to evaluate agreement across all categories without merging, ensuring a detailed understanding of LLM performance in distinguishing between fine-grained classifications. Using bootstrap resampling with 1000 iterations, the 95 % confidence interval ranges (CI) were reported. Additionally, Cohen's κ was calculated for binary classifications of CAD-RADS scores, grouping values as 0 to 2 and, CAD-RAD 3 or greater, as a CAD-RADS score of 3 or greater would typically prompt further patient evaluation. For multi-class tasks, true and false positives were reported as proportions, while binary tasks used 2 × 2 tables to calculate sensitivity, specificity, precision, recall, and subsequently F1 scores. A detailed description on these performances, and their calculations is provided in the supplements. Classification



Case 1

Coronary Calcium scoring
The total Agatston Score is 49.

Coronary Angiography CT
Supply type: Right-dominant

Image quality/assessability: Good image quality

LM: Normal origin from the left coronary sinus. No evidence of stenosis, and no plaque visible.

LAD: Calcified plaques with resulting 20% stenosis in the first diagonal branch. Otherwise, no stenosis or no plaques.

LCX: No evidence of relevant plaque, no stenosis.

RCA: Originating from the right coronary sinus. No evidence of relevant plaque and no stenosis.

Additional findings: Pericentral pulmonary embolism, as well as peripheral emboli. No evidence of right heart strain. No pleural or pericardial effusion. Bones without pathological changes.

A discussion of the findings was held, and the patient was transferred to the emergency department due to the pulmonary embolism.

Case 2

Coronary Calcium scoring
The total Agatston Score is 210.

Coronary Angiography CT
Supply type: Right-dominant

Image quality/assessability: Diagnostic image quality in all segments

LM: Normal origin from the left coronary sinus. No evidence of stenosis, no plaque.

LAD: Mixed-calcified plaques with resulting 75% stenosis in the proximal segment. No further stenosis.

LCX: No evidence of relevant plaque and no stenosis.

RCA: Originating from the right coronary sinus. In the mid-segment, there is mild stenosis of approximately 40% due to calcified plaques.

In the additionally performed CT-FFR, a value of 0.71 was noted for the LAD.

Additional findings: Mild atherosclerotic changes of the aorta. No pleural or pericardial effusion. Lymph nodes are of normal size. Upper abdomen, as far as visualized, is unremarkable. Vertebral hemangioma in T11.

Fig. 1. A dataset of 200 cardiac CT reports was created and used to assess the performance of seven Large Language Models: GPT-3.5, GPT-4o, Mistral 7b, Mixtral 8 × 7b, Llama 3 8b, Llama 3 8b with a 64k context length, and Llama 3 70b. The accuracy of each model's generated CAD-RADS 2.0 score using in-context learning was compared to a ground truth established by two board-certified cardiothoracic radiologists.

Table 1
Characteristics of each used LLM.

Model	Manufacturer	Open-Source	GPU Size (GB)	Context Window	Brief Description
GPT-4o	OpenAI	No	N/A	32k	Advanced language model by OpenAI, known for high performance across various tasks
GPT-3.5-Turbo	OpenAI	No	N/A	16k	Efficient and cost-effective model by OpenAI, widely used for various applications
Llama 3.1 70B 4Q	Meta	Yes	~24	4096	Large quantized version of Meta's Llama 3.1, offering high performance with reduced precision
Llama 3.1 8B 6Q	Meta	Yes	~24	4096	Smaller quantized version of Llama 3.1, balancing performance and resource requirements
Llama 3.1 6Q (64k context)	Meta	Yes	~24	65536	Extended context version of Llama 3.1, allowing for processing of longer sequences
Mistral 7B	Mistral.ai	Yes	~24	8192	Efficient open-source model, suitable for various NLP tasks
Mixtral 8 × 7B	Mistral.ai	Yes	~48	32768	Sparse Mixture of Experts model, offering high performance across multiple benchmarks

Abbreviations: GPU, Graphics Processing Unit; GB, Gigabyte; NLP, Natural Language Processing; 4Q, 6Q, Quantized versions with different bit precision; B, Billion (parameter count in models); k, Thousand (context window size).

performance of the LLMs was measured using the F1 score for binary tasks, and the proportion of correct interpretations, scaled from 0 to 1, for multi-class categories.

3. Results

3.1. *Characteristics*

In the final set of 200 cases, the Agatston score ranged from 0 to 6024, with a median score of 63 [IQR 16–151]. A total of 112 cases (56 %) had a CAD-RADS score of 2 or lower, indicating minimal to mild stenosis, whereas 10 cases (5 %) had total occlusion (CAD-RADS 5). Further details, including plaque burden, ischemia, and other modifiers, are displayed in Table 2.

3.2. *Performance of LLMs in evaluating CCTA reports*

3.2.1. *CAD-RADS score*

For CAD-RADS score classification, GPT-4o and Llama 3 70B achieved the highest performance, correctly classifying 198 out of 200 cases (99 %). Llama 3 8B closely followed, with 196 correct classifications (98 %), while GPT-3.5 and Mixtral 8 × 7B each correctly classified 194 cases (97 %). Mistral 7B performed slightly inferior, correctly classifying 191 cases (95.5 %). The lowest performance was observed with Llama 3 8B with a 64k context length, which correctly classified 177 cases (88.5 %). Fig. 2 provides the distribution of CAD-RADS categories among distinct LLMs, while Table 3 displays the counts in comparison to the ground truth. A confusion matrix on the performance of each LLM and detailed contingencies, including the performances of precision, recall and F1-Scores, including their respected 95 % CI ranges regarding the CAD-RADS classification are reported in the Supplements.

Table 2
Characteristics of the cases described in synthetic reports.

Characteristics	CCTA Reports (*n=200*)
CAD RADS	
0	49 (24.5 %)
1	34 (17.0 %)
2	29 (14.5 %)
3	26 (13.0 %)
4A	32 (16.0 %)
4B	15 (7.5 %)
5	10 (5.0 %)
N	5 (2.5 %)
Coronary artery calcification	
Agatston Score[a]	63 [16–151]
Agatston Score, range	0–6024
Plaque	
Absence of calcification or plaque	38 (19.0 %)
P1 (mid)	97 (48.5 %)
P2 (moderate)	35 (17.5 %)
P3 (severe)	19 (9.5 %)
P4 (extensive)	11 (5.5 %)
High-Risk-Plaques (HRP)	24 (12 %)
I – Modifier	
I +	13 (6.5 %)
I +/−	4 (2.0 %)
I –	13 (6.5 %)
Stent (S–modifier)	6 (3.0 %)
Bypass-graft (G–modifier)	5 (2.5 %)
Exception (E–modifier)	7 (3.5 %)

Abbreviations: *CAD-RADS,* Coronary Artery Disease Reporting and Data System, *P,* Plaques; *I,* Ischemia.

[a] Data is presented with median and interquartile range in square brackets.

3.2.2. *Plaque burden*

For assignment of coronary plaque, GPT-4o and Llama 3 70B achieved the highest accuracy, correctly classifying 198 out of 200 cases (99 %), respectively. Mixtral 8 × 7B closely followed with 196 correct classifications (98 %). Llama 3 8B correctly classified 98 cases (49 %) and Llama 3 8B with a 64k context length only 46 cases (23 %). GPT-3.5 and Mistral 7B showed poor performance with only 37 (18.5 %) and 38 (19 %) correct classifications, respectively. Details are displayed in Table 4, while Fig. 3 provides the distribution of P-categories across LLMs. Interrater reliability between each LLM and the ground truth, assessed by Krippendorff's alpha and Cohen's kappa for CAD-RADS, as well as plaque quantification assessment, is summarized in Table 5. A confusion matrix on the performance of each LLM and detailed contingencies, and performances of precision, recall and F1-Scores, including their respected 95 % CI ranges regarding the plaque burden classification are reported in the Supplements.

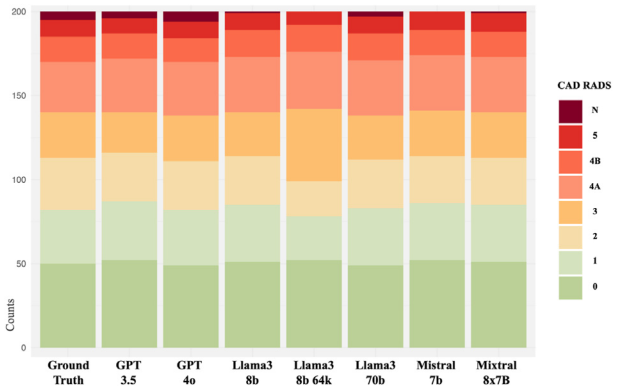


Fig. 2. Stacked bar chart comparing CAD-RADS 2.0 scores by different large language models (GPT-3.5, GPT-4o, Llama 3 variants, and Mixtral 8 × 7B) against the reference standard.

Table 3a
Comparison of Stenosis Severity assignment into CAD-RADS by Large Language Models.

CAD-RADS	Ground Truth	GPT 3.5	GPT 4o	Llama 3 8b	Llama 3 8b 64k	Llama 3 70b	Mistral 7b	Mixtral-8x7B
0	50	52	49	51	52	49	52	51
1	32	35	33	34	26	34	34	34
2	31	29	29	29	21	29	28	28
3	27	24	27	26	43	26	27	27
4A	30	32	32	33	34	33	33	33
4B	15	15	14	16	16	16	15	15
5	10	9	10	10	8	10	11	11
N	5	4	6	1	0	3	0	1
Correctly Classified	N/A	194/200 (97 %)	198/200 (99 %)	196/200 (98 %)	177/200 (88.5 %)	198/200 (99 %)	191/200 (95.5 %)	194/200 (97 %)

Number of cases classified by various large language models and their variants into CAD-RADS categories. Each cell shows the count of cases assigned to the respective category, which ranges from 0 (no stenosis), to 5 (total coronary occlusion).
Abbreviations: *CAD-RADS,* Coronary Artery Disease Reporting and Data System; *N*, non-diagnostic; *N/A,* not applicable.

Table 3b
F1-Scores by Large Language Models in CAD-RADS classification.

CAD-RADS	GPT–3.5	GPT–4o	Llama 3 8b	Llama 3 8b 64k	Llama 3 70b	Mistral 7b	Mixtral 8 × 7B
0	0.97 [0.932, 1.0]	0.99 [0.965, 1.0]	0.98 [0.949, 1.0]	0.97 [0.928, 1.0]	1.0 [1.0, 1.0]	0.97 [0.928, 1.0]	0.98 [0.949, 1.0]
1	0.986 [0.951, 1.0]	1.0 [1.0, 1.0]	1.0 [1.0, 1.0]	0.867 [0.757, 0.947]	1.0 [1.0, 1.0]	1.0 [1.0, 1.0]	1.0 [1.0, 1.0]
2	0.966 [0.909, 1.0]	1.0 [1.0, 1.0]	1.0 [1.0, 1.0]	0.84 [0.718, 0.933]	1.0 [1.0, 1.0]	0.982 [0.939, 1.0]	0.982 [0.941, 1.0]
3	0.96 [0.894, 1.0]	0.981 [0.935, 1.0]	1.0 [1.0, 1.0]	0.754 [0.625, 0.857]	1.0 [1.0, 1.0]	0.943 [0.864, 1.0]	0.981 [0.938, 1.0]
4A	1.0 [1.0, 1.0]	0.985 [0.949, 1.0]	0.985 [0.941, 1.0]	0.97 [0.92, 1.0]	0.985 [0.952, 1.0]	0.985 [0.949, 1.0]	0.985 [0.945, 1.0]
4B	0.933 [0.8, 1.0]	0.966 [0.865, 1.0]	0.968 [0.88, 1.0]	0.968 [0.897, 1.0]	0.968 [0.882, 1.0]	0.867 [0.71, 0.973]	0.933 [0.812, 1.0]
5	0.947 [0.8, 1.0]	1.0 [1.0, 1.0]	1.0 [1.0, 1.0]	0.889 [0.667, 1.0]	1.0 [1.0, 1.0]	0.952 [0.8, 1.0]	0.952 [0.833, 1.0]
N	0.889 [0.4, 1.0]	1.0 [1.0, 1.0]	0.333 [0.0, 0.8]	0.0 [0.0, 0.0]	0.75 [0.0, 1.0]	0.0 [0.0, 0.0]	0.333 [0.0, 0.8]

Note.–Data in square brackets represent the 95 % confidence interval range.

Table 4a
Comparison of Plaque Burden assignment into P-modifiers by Large Language Models.

P Modifier	Ground Truth	GPT 3.5	GPT 4o	Llama 3 8b	Llama 3 8b 64k	Llama 3 70b	Mistral 7b	Mixtral-8x7B
P0	38	200	38	139	191	38	199	41
P1	97	0	97	43	6	97	0	96
P2	35	0	35	15	3	35	1	34
P3	19	0	19	3	0	19	0	18
P4	11	0	11	0	0	11	0	11
Correctly Classified	N/A	38/200 (19 %)	198/200 (99 %)	98/200 (49 %)	46/200 (23 %)	198/200 (99 %)	38/200 (19 %)	196/200 (98 %)

Number of cases classified by various large language models and their variants into P-modifiers according to CAD-RADS 2.0. Each cell shows the count of cases assigned to the respective P-modifier category, which ranges from 0 (no plaques) to P4 (extensive plaque burden).
Abbreviations: *N/A*, not applicable.

Table 4b
F1-Scores by Large Language Models in Plaque-burden classification.

P Modifier	GPT–3.5	GPT–4o	Llama 3 8b Q6	Llama 3 8b 64k	Llama 3 70b	Mistral 7b	Mixtral
P0	0.314 [0.239, 0.389]	1.0 [1.0, 1.0]	0.423 [0.321, 0.511]	0.326 [0.249, 0.41]	1.0 [1.0, 1.0]	0.315 [0.233, 0.389]	0.961 [0.909, 1.0]
P1	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]	0.614 [0.516, 0.707]	0.117 [0.025, 0.198]	1.0 [1.0, 1.0]	0.0 [0.0, 0.0]	0.995 [0.984, 1.0]
P2	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]	0.6 [0.418, 0.746]	0.158 [0.0, 0.303]	1.0 [1.0, 1.0]	0.056 [0.0, 0.167]	0.986 [0.951, 1.0]
P3	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]	0.273 [0.0, 0.519]	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]	0.0 [0.0, 0.0]	0.973 [0.895, 1.0]
P4	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]	0.0 [0.0, 0.0]	1.0 [1.0, 1.0]

Note.–Data in square brackets represent the 95 % confidence interval range.

3.2.3. *Other modifiers (HRP, I, S, G and E)*

For *HRP* identification, GPT-4o correctly classified 23 out of 23 cases (100 %), while Llama 3 70B attributed correct scores to 21 out of 23 cases (92.3 %). In contrast, Mistral 7B, Mistral 8 × 7B, and Llama 3 8B failed to identify any HRP cases (0 %). For the I-modifier, GPT-4o achieved the highest performance with 17 true positives and 0 false negatives, while GPT-3.5 had 7 true positives and 9 false negatives. Llama 3 70B identified 15 true positives with 1 false negative. For bypass graft (*G-modifier*) detection, GPT-3.5, GPT-4o, and Mistral 7B correctly identified all 5 cases (100 %), while Mistral 8 × 7B identified 2 out of 5 cases (40 %). In terms of stent identification (*S-modifier*), GPT-3.5, GPT-4o, and Llama 3 70B correctly identified 6 out of 6 cases (100 %), whereas Mixtral 8 × 7B identified 2 out of 6 cases (33.3 %). Notably, GPT-3.5 falsely assigned 23 cases as having stents. Further details on model performance are shown in Table 6.

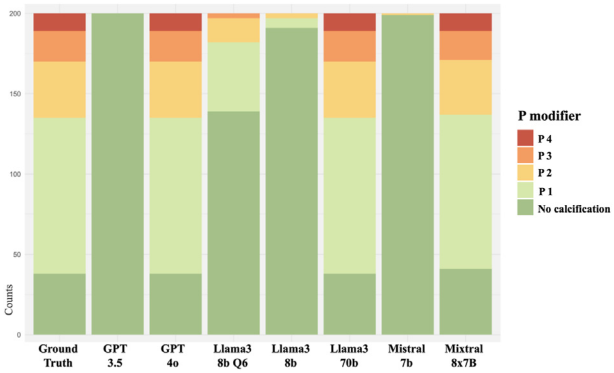


Fig. 3. Stacked bar chart comparing P-modifier assignments by different large language models (GPT-3.5, GPT-4o, Llama 3 variants, and Mixtral 8 × 7B). Note how GPT-3.5, Llama 3 8b, and Mistral 7b, wrongfully assumed no calcifications or plaque for most cases.

3.2.4. Accuracy for comprehensive CAD RADS 2.0 scoring including all modifiers

GPT-4o achieved the highest accuracy by correctly attributing the whole CAD-RADS 2.0 sequence in 186 out of 200 cases (93.0 %), closely followed by Llama 3 70B that correctly classified 185 out of 200 cases (92.5 %). Mixtral 8 × 7B demonstrated intermediate performance, correctly classifying 156 out of 200 cases (78.0 %). GPT-3.5, Llama 3 8B, and Llama 3 8B with 64k context length each correctly classified only 32 out of 200 cases (16.0 %). A summary of the performance in each distinct aspect of the CAD-RASD 2.0 classification is visualized in the radar plots (Fig. 4).

4. Discussion

Our study aimed to evaluate the capability of state-of-the-art LLMs with in-context learning to autonomously generate CAD-RADS 2.0 scores from cardiac CT reports. The main findings of our study are as follows: i) all LLMs performed reasonably well in categorizing stenosis according to CAD-RADS categories, ii) for assignment of plaque burden and other modifiers, such as stent assessment, a substantial difference was present, and iii) upon assessing all modifiers and thus complete CAD-RADS 2.0 score, which would guide clinical decision making, only GPT-4o and Llama 3 70B demonstrated excellent overall accuracy, achieving 93 % and 92.5 %, respectively.

A recent study by Monroe et al. evaluated 30 cardiac imaging-related questions across three separate chat sessions, with GPT-3.5 providing 61 % correct answers and GPT-4 75 %. However, misleading answers were observed in 39 % of GPT-3.5's responses and 29 % of GPT-4's.[26] On the one hand, the increase of chatbot performance by inclusion of specialized knowledge is known and significantly improved imaging recommendations of pretrained LLM.[20] On the other hand, we observed substantially better performance for the more recent LLMs. The progress form GPT-3.5 to GPT-4o in radiological tasks is well documented, with improved performances in zero-shot extraction of oncological information from

Table 5
Inter-rater reliability for CAD-RADS and Plaque evaluation across large language models.

	GPT 3.5	GPT 4o	Llama 3 8b	Llama 3 8b 64k	Llama 3 70b	Mistral 7b	Mixtral 8 × 7B
CAD RADS score							
Krippendorff's α	0.96 (0.93, 0.99)	0.98 (0.95, 1.00)	0.95 (0.88, 1.00)	0.90 (0.81, 0.97)	1.00 (0.99, 1.00)	0.91 (0.83, 0.99)	0.97 (0.92, 1.00)
Cohen's κ[a]	0.96 (0.92, 0.99)	0.99 (0.97, 1.00)	0.98 (0.95, 0.99)	0.81 (0.73, 0.89)	1.00	0.96 (0.92, 0.99)	0.99 (0.97, 1.00)
Plaque							
Krippendorff's α	−0.64 (−0.70, −0.57)	1.00	−0.09 (−0.24, 0.08)	−0.53 (−0.62, −0.44)	1.00	−0.62 (−0.69, −0.54)	0.96 (0.90, 1.00)
Cohen's κ[a]	0.00 (−0.14, 0.14)	1.00 (1.00, 1.00)	0.18 (0.05, 0.32)	0.02 (-0.12, 0.16)	1.00 (1.00, 1.00)	0.00 (-0.14, 0.14)	0.95 (0.91, 0.99)

Note.—Data in parenthesis represent the 95 % confidence interval range.

[a] The κ value for CAD-RADS was calculated for scores 0–2 versus ≥3, and for plaque as absence versus P1–P4.

Table 6
Performance of Modifier assignment by Large Language Models.

Modifier	Diagnostic Metric	GPT 3.5	GPT 4o	Llama 3 8b	Llama 3 8b 64K	Llama 3 70b	Mistral 7b	Mixtral 8 × 7B
N	True Positive Rates	5/13 (33.3 %)	13/13 (100 %)	1/13 (7.7 %)	0/13 (0 %)	12/13 (92.3 %)	1/13 (7.7 %)	12/13 (92.3 %)
	False Positives (*n*)	2	1	0	0	1	0	0
HRP	True Positive Rates	10/23 (43.5 %)	23/23 (100 %)	0/23 (0.0 %)	2/23 (8.7 %)	21/13 (92.3 %)	0/23 (0.0 %)	0/23 (0.0 %)
	False Positives (*n*)	2	9	0	0	7	0	0
I + or I ±	True Positive Rates	7/17 (41.2 %)	17/17 (100 %)	8/17 (47.1 %)	1/17 (5.9 %)	15/17 (88.2 %)	12/17 (70.6 %)	15/17 (88.2 %)
	False Positives (*n*)	0	1	0	0	1	2	0
I-	True Positive Rates	3/13 (23.1 %)	13/13 (100 %)	5/13 (38.5 %)	0/13 (0 %)	13/13 (100 %)	2/13 (15.4 %)	5/13 (38.5 %)
	False Positives (*n*)	0	0	0	0	0	0	0
G	True Positive Rates	5/5 (100 %)	5/5 (100 %)	0/5 (0.0 %)	0/5 (0.0 %)	0/5 (0.0 %)	5/5 (100 %)	2/5 (40.0 %)
	False Positives (*n*)	0	0	0	0	0	0	0
S	True Positive counts	6/6 (100 %)	6/6 (100 %)	0/6 (0.0 %)	0/6 (0.0 %)	6/6 (100 %)	0/6 (100 %)	2/6 (33.3 %)
	False Positives (*n*)	23	0	0	0	0	0	1
E	True Positive Rates	5/7 (71.4 %)	7/7 (100 %)	1/7 (14.3 %)	0/7 (0 %)	7/7 (100 %)	0/7 (0 %)	7/7 (100 %)
	False Positives (*n*)	11	2	0	0	1	0	1

Abbreviations: *N*, non-diagnostic; *HPR*, High-Risk Plaque; *I*, Ischemia – with *I+* and *I-* indicating presence, or absence of ischemia, respectively. *I+/−*, inconclusive; *G*, bypass graft; *S*, stent; *E*, exception.

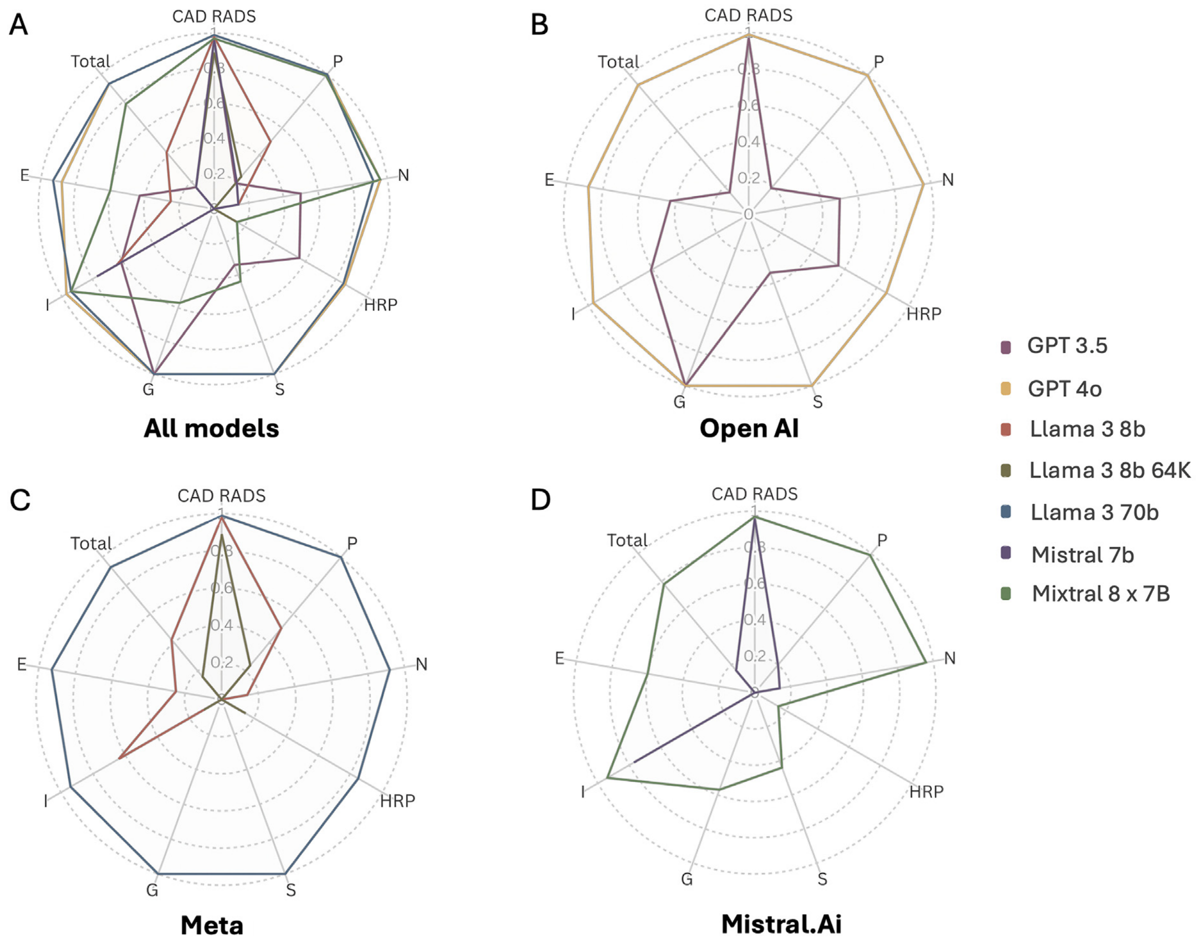


Fig. 4. Radar Plot displaying the diagnostic performance of each large language model (LLM) in analyzing cardiac CT reports according to the updated CAD-RADS 2.0 classification. Each distinct category is displayed, as well as the entire scoring with correct sequence. A) displays all tested LLMs, B) displays the radar plot of GPT-3.5, GPT-4o (Open.AI), C) of Llama 3 8b, Llama 3 8b with a 64k context length, and Llama 3 70b (Meta), and D) of Mistral 7b, Mixtral 8 × 7b (mistral.ai).

clinical progress notes[27] simplifying and translating radiology reports,[28] or even in various radiological board-certification examinations.[29,30] In a comparison of commercial and open-source LLMs for labeling chest radiograph reports, GPT-4o achieved slightly higher F1 scores than open-source models few-shot prompting narrowed the performance gap, with the ensemble model matching GPT-4o on the institutional dataset.[31] These results are confirmatory to our study, as we observe only slight differences, when few-shot prompting is performed.

We noted a markedly worse performance in assessment of P-modifiers, especially for older LLMs. We attribute that to the fact that these models with fixed information base possess an inherent inability to incorporate recent research results, their inability to check for errors and lack of design for specialized medical reasoning.[32] On the other hand, the partly vague nature of the CAD-RADS 2.0 guidelines with regard to plaque quantification must be acknowledged, as here multiple possibilities, such as a visual grading, semiquantitative approaches (SIS score) and quantitative approaches involving the Agatston-Score are simultaneously proposed, and the most severe one should be reported.[23] Here, standardizing plaque quantification and reporting, with focusing on plaque composition and integrating the SIS in a cardiac CT report would be most favorable and necessary.[33] This is crucial for facilitating large-scale, structured reporting initiatives.[34] With further refinement, reporting standards and proposed by radiological guidelines could also be tested for their applicability and potential for successful integration into large scale structured reporting based on the performance of task-specific LLMs.

A major concern regarding the incorporation of LLMs into medical decision making is their lack of trustworthiness, and aptitude to produce vague or conjectural answers, commonly referred as ‘hallucinating’.[35,36] This was particularly evident in our study with GPT-3.5, which incorrectly assumed the presence of stents (S-modifier) in 23 cases. In contrast, more recent LLMs exhibited this behavior only to a minor extent. This is confirmatory to a recent study analyzing simple CAD RADS 1.0 scores in 100 reports, where GPT-3.5 through hallucination even suggested CAD-RADS 6 categories.[37]

While GPT.4o overall performed best, its cloud-based processing hampers the integration into current clinical workflows due to data privacy concerns.[35] The transmission of sensitive patient data with commercial companies and upload of data to external servers, poses a significant data security issue.[38] In contrast, locally hosted models (such as Llama 3 70B), which performed competitively well, allow for a more secure solution by keeping data within on-premises servers within enclosed institutional infrastructure. In general, open-source models, which can be deployed locally or within private cloud environments, reduce reliance on external servers, thus minimizing data exposure and third-party involvement. However, local deployment requires robust cybersecurity measures, as hospital and lab infrastructures may still be vulnerable to attacks.[39] Conversely, proprietary cloud-based models often incorporate strong security frameworks, with some vendors offering HIPAA-compliant solutions. Recent advancements have introduced frameworks like LocalAI (https://localai.io) and Ollama (https://ollama.com), which enable institutions to deploy LLMs on-premises, preserving data privacy while maintaining processing efficiency. LocalAI functions as an OpenAI-compatible endpoint, allowing for seamless integration of Hugging Face-hosted models (https://huggingface.co), while Ollama provides a user-friendly deployment interface suited for both research and clinical applications. However, efficient inference—especially for large models—requires high-performance GPUs with substantial VRAM, highlighting the trade-off between computational demand and data security. These considerations are critical for translating LLMs into real-world clinical workflows, where balancing accuracy, security, and infrastructure feasibility will dictate their practical utility in radiological structured reporting tasks, such as automated CAD-RADS 2.0 classification.

Our study has some limitations: Firstly, to ensure data privacy and avoid ethical concerns, this study was conducted using synthetically generated cardiac CT reports. Further research is needed to validate LLM performance in reports from clinical routine. Furthermore, the reports were written in German language and as some LLMs are primarily trained for English language, it cannot be entirely excluded that our observed performances represent an underestimation.[40] This study focused on LLM accuracy for CAD-RADS 2.0 and did not assess potential time savings. Evaluating these aspects requires a dedicated study comparing LLM vs. expert classifications in real-world settings. Lastly, while our study focuses on CCTA reports and CAD-RADS 2.0, the generalizability of these models to other radiological structured reporting systems, such as Lung-RADS or Breast Imaging-RADS, remains unclear.

5. Conclusion

To conclude, in-context learning-enabled LLMs, such as GPT-4o and Llama 3 70B, accurately generate CAD-RADS 2.0 scores from cardiac CT reports, potentially enhancing reporting consistency and efficiency. Open-source models offer a secure alternative to cloud-based solutions, addressing data privacy concerns while maintaining high performance. Future validation on a diverse set of real-world data is essential for clinical integration.

Funding

German Research Foundation (DFG) - SFB 1597–499552394.

Declaration of competing interest

AVS Unrestricted research grant, speakers' bureau (Siemens Healthineers), Consultant, shareholder (Elucid Bioimaging).

TE Unrestricted research grant, speakers' bureau (Siemens Healthineers), Consultant Circle Cardiovascular Imaging.

FB Unrestricted research grant, speakers' bureau (Siemens Healthineers, Bayer Healthcare).

MTH Speakers' bureau (Siemens Healthineers).

All other authors declare no potential conflicts of interest.

Acknowledgments

P.A.: Hans A. Krebs Medical Scientist Program (Uniklinikum Freiburg). AR was supported by the Berta-Ottenstein-Programme for Clinician Scientists, Faculty of Medicine, University of Freiburg.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcct.2025.03.007.

References

1. Vaduganathan M, Mensah GA, Turco JV, Fuster V, Roth GA. The Global burden of cardiovascular diseases and risk. *J Am Coll Cardiol.* 2022;80(25):2361–2371. https://doi.org/10.1016/j.jacc.2022.11.005.
2. Gulati M, Levy PD, Mukherjee D, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR guideline for the evaluation and diagnosis of chest Pain: a report of the American College of Cardiology/American Heart Association Joint Committee on clinical practice guidelines. *Circulation.* 2021;144(22):e368–e454. https://doi.org/10.1161/CIR.0000000000001029.
3. Vrints C, Andreotti F, Koskinas KC, et al. 2024 ESC Guidelines for the management of chronic coronary syndromes: developed by the task force for the management of chronic coronary syndromes of the European Society of Cardiology (ESC) Endorsed by the European Association for Cardio-Thoracic Surgery (EACTS). *Eur Heart J.* Published online August 30, 2024:ehae177. https://doi.org/10.1093/eurheartj/ehae177.
4. Cury RC, Abbara S, Achenbach S, et al. CAD-RADS(TM) coronary artery disease - reporting and data system. An expert consensus document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Radiology (ACR) and the North American Society for Cardiovascular Imaging (NASCI). Endorsed by the American College of Cardiology. *J Cardiovasc Comput Tomogr.* 2016;10(4):269–281. https://doi.org/10.1016/j.jcct.2016.04.005.
5. Ferencik M, Mayrhofer T, Bittner DO, et al. Use of high-risk coronary atherosclerotic plaque detection for risk stratification of patients with stable chest pain: a secondary analysis of the PROMISE randomized clinical trial. *JAMA Cardiol.* 2018;3(2):144–152. https://doi.org/10.1001/jamacardio.2017.4973.
6. Pontone G, Rossi A, Baggiano A, et al. Progression of non-obstructive coronary plaque: a practical CCTA-based risk score from the PARADIGM registry. *Eur Radiol.* 2024;34(4):2665–2676. https://doi.org/10.1007/s00330-023-09880-x.
7. Bittencourt MS, Hulten E, Ghoshhajra B, et al. Prognostic value of nonobstructive and obstructive coronary artery disease detected by coronary computed tomography angiography to identify cardiovascular events. *Circ Cardiovasc Imag.* 2014;7(2):282–291. https://doi.org/10.1161/CIRCIMAGING.113.001047.
8. Cury RC, Leipsic J, Abbara S, et al. CAD-RADS^{TM} 2.0 – 2022 coronary artery disease – reporting and data system an expert consensus document of the Society of Cardiovascular Computed Tomography (SCCT), the American College of Cardiology (ACC), the American College of Radiology (ACR) and the North America Society of Cardiovascular Imaging (NASCI). *Radiol Cardiothorac Imag.* 2022;4(5):e220183. https://doi.org/10.1148/ryct.220183.
9. Huang Z, Yang Y, Wang Z, et al. Comparison of prognostic value between CAD-RADS 1.0 and CAD-RADS 2.0 evaluated by convolutional neural networks based CCTA. *Heliyon.* 2023;9(5):e15988. https://doi.org/10.1016/j.heliyon.2023.e15988.
10. Ramanathan S, Al Heidous M, Alkuwari M. Coronary artery disease-reporting and data system (CAD-RADS): strengths and limitations. *Clin Radiol.* 2019;74(6):411–417. https://doi.org/10.1016/j.crad.2019.01.003.
11. Catapano F, Moser LJ, Francone M, et al. Competence of radiologists in cardiac CT and MR imaging in Europe: insights from the ESCR Registry. *Eur Radiol.* 2024;34(9):5666–5677. https://doi.org/10.1007/s00330-024-10644-4.
12. Can E, Uller W, Vogt K, et al. Large Language models for simplified interventional radiology reports: a comparative analysis. *Acad Radiol.* 2024. https://doi.org/10.1016/j.acra.2024.09.041. Published online September 30.
13. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology.* 2023;307(4). https://doi.org/10.1148/radiol.230725.
14. Mitsuyama Y, Tatekawa H, Takita H, et al. Comparative analysis of GPT-4-based ChatGPT's diagnostic performance with radiologists using real-world radiology reports of brain tumors. *Eur Radiol.* 2024. https://doi.org/10.1007/s00330-024-11032-8. Published online.
15. Doshi R, Amin KS, Khosla P, Bajaj SS, Chheang S, Forman HP. Quantitative evaluation of Large Language Models to streamline radiology report impressions: a multimodal retrospective analysis. *Radiology.* 2024;310(3):e231593. https://doi.org/10.1148/radiol.231593.
16. Hagar MT, Soschynski M, Benndorf M, et al. Enhancing radiation dose efficiency in prospective ECG-triggered coronary CT angiography using calcium-scoring CT. *Diagnostics.* 2023;13(12):2062. https://doi.org/10.3390/diagnostics13122062.
17. Hoffmann U, Ferencik M, Udelson JE, et al. Prognostic value of noninvasive cardiovascular testing in patients with stable chest pain. *Circulation.* 2017;135(24):2320–2332. https://doi.org/10.1161/CIRCULATIONAHA.116.024360.
18. SCOT-HEART Investigators, Newby DE, Adamson PD, et al. Coronary CT angiography and 5-year risk of myocardial infarction. *N Engl J Med.* 2018;379(10):924–933. https://doi.org/10.1056/NEJMoa1805971.
19. Busch F, Prucker P, Komenda A, et al. Multilingual feasibility of GPT-4o for automated Voice-to-Text CT and MRI report transcription. *Eur J Radiol.* 2025;182:111827. https://doi.org/10.1016/j.ejrad.2024.111827.
20. Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR appropriateness guidelines. *Radiology.* 2023;308(1):e230970. https://doi.org/10.1148/radiol.230970.
21. Meddeb A, Lüken S, Busch F, et al. Large Language model ability to translate CT and MRI free-text radiology reports into multiple languages. *Radiology.* 2024;313(3):e241736. https://doi.org/10.1148/radiol.241736.

22. Busch F, Hoffmann L, dos Santos DP, et al. Large language models for structured reporting in radiology: past, present, and future. *Eur Radiol*. 2024. https://doi.org/10.1007/s00330-024-11107-6. Published online October 23.
23. Cury RC, Leipsic J, Abbara S, et al. CAD-RADS^{TM} 2.0 - 2022 coronary artery disease-reporting and data system: an expert consensus document of the society of Cardiovascular Computed Tomography (SCCT), the American College of Cardiology (ACC), the American College of Radiology (ACR), and the North America Society of Cardiovascular Imaging (NASCI). *J Cardiovas Comp Tomograp*. 2022;16(6):536–557. https://doi.org/10.1016/j.jcct.2022.07.002.
24. Williams MC, Moss AJ, Dweck M, et al. Coronary artery plaque characteristics associated with adverse outcomes in the SCOT-HEART study. *J Am Coll Cardiol*. 2019;73(3):291–301. https://doi.org/10.1016/j.jacc.2018.10.066.
25. Maurovich-Horvat P, Hoffmann U, Vorpahl M, Nakano M, Virmani R, Alkadhi H. The napkin-ring sign: CT signature of high-risk coronary plaques? *JACC Cardiovasc Imag*. 2010;3(4):440–444. https://doi.org/10.1016/j.jcmg.2010.02.003.
26. Monroe CL, Abdelhafez YG, Atsina K, Aman E, Nardo L, Madani MH. Evaluation of responses to cardiac imaging questions by the artificial intelligence large language model ChatGPT. *Clin Imag*. 2024;112:110193. https://doi.org/10.1016/j.clinimag.2024.110193.
27. Sushil M, Kennedy VE, Mandair D, Miao BY, Zack T, Butte AJ. CORAL: expert-curated oncology reports to advance language model inference. *NEJM AI*. 2024;1(4): AIdbp2300110. https://doi.org/10.1056/AIdbp2300110.
28. Lyu Q, Tan J, Zapadka ME, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Vis Comput Ind Biomed Art*. 2023;6:9. https://doi.org/10.1186/s42492-023-00136-5.
29. R B, S K, Rr B. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5). https://doi.org/10.1148/radiol.230582.
30. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198.
31. Dorfner FJ, Jürgensen L, Donle L, et al. Comparing commercial and open-source Large Language Models for labeling chest radiograph reports. *Radiology*. 2024; 313(1):e241139. https://doi.org/10.1148/radiol.241139.
32. Srivastav S, Chandrakar R, Gupta S, et al. ChatGPT in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis. Cureus. 15(7): e41435. doi:10.7759/cureus.41435.
33. Shaw LJ, Blankstein R, Bax JJ, et al. Society of cardiovascular computed tomography/North American Society of cardiovascular imaging – expert consensus document on coronary CT imaging of atherosclerotic plaque. *J Cardiovas Comp Tomograp*. 2021;15(2):93–109. https://doi.org/10.1016/j.jcct.2020.11.002.
34. Jorg T, Halfmann MC, Arnhold G, et al. Implementation of structured reporting in clinical routine: a review of 7 years of institutional experience. *Insights Imag*. 2023; 14(1):61. https://doi.org/10.1186/s13244-023-01408-7.
35. Shen Y, Heacock L, Elias J, et al. ChatGPT and other Large Language Models are double-edged swords. *Radiology*. 2023;307(2):e230163. https://doi.org/10.1148/radiol.230163.
36. Xiao Y, Wang WY. *On Hallucination and Predictive Uncertainty in Conditional Language Generation*. 2021. https://doi.org/10.48550/arXiv.2103.15025. Published online March 28.
37. Silbergleit M, Tóth A, Chamberlin JH, et al. ChatGPT vs gemini: comparative accuracy and efficiency in CAD-RADS score assignment from radiology reports. *J Imaging Inform Med*. 2024. https://doi.org/10.1007/s10278-024-01328-y. Published online.
38. Mehrtak M, SeyedAlinaghi S, MohsseniPour M, et al. Security challenges and solutions using healthcare cloud computing. *J Med Life*. 2021;14(4):448–461. https://doi.org/10.25122/jml-2021-0100.
39. Savage CH, Kanhere A, Parekh V, et al. Open-source Large Language Models in radiology: a review and tutorial for practical research and clinical deployment. *Radiology*. 2025;314(1):e241073. https://doi.org/10.1148/radiol.241073.
40. Cozzi A, Pinker K, Hidber A, et al. BI-RADS category assignments by GPT-3.5, GPT-4, and google bard: a multilanguage study. *Radiology*. 2024;311(1):e232133. https://doi.org/10.1148/radiol.232133.

Glossary of Abbreviations

AI: Artificial intelligence
CAD: Coronary artery disease
CAD-RADS: Coronary artery disease-reporting and data system
CCTA: Coronary CT angiography
CI: Confidence interval
FFR-CT: Fractional flow reserve derived from CT
HRP: High-risk plaque
IQR: Interquartile range
LLMs: Large language models
NLP: Natural language processing
SD: Standard deviation