Original Article | Cardiovascular Imaging

eISSN 2005-8330
https://doi.org/10.3348/kjr.2025.0293
Korean J Radiol 2025;26(9):817-831

Korean Journal of Radiology
KJR

Large Language Models for CAD-RADS 2.0 Extraction From Semi-Structured Coronary CT Angiography Reports: A Multi-Institutional Study

Dabin Min[1,2*], Kwang Nam Jin[3,4*], SangHeum Bang[3], Moon Young Kim[3,4], Hack-Lyoung Kim[5], Won Gi Jeong[6], Hye-Jeong Lee[7], Kyongmin Sarah Beck[8], Sung Ho Hwang[9], Eun Young Kim[10], Chang Min Park[1,4,11,12]

[1]Interdisciplinary Program in Bioengineering, Seoul National University Graduate School, Seoul, Republic of Korea
[2]Integrated Major in Innovative Medical Science, Seoul National University Graduate School, Seoul, Republic of Korea
[3]Department of Radiology, Boramae Medical Center, Seoul National University College of Medicine, Seoul, Republic of Korea
[4]Seoul National University College of Medicine, Seoul, Republic of Korea
[5]Division of Cardiology, Department of Internal Medicine, Boramae Medical Center, Seoul National University College of Medicine, Seoul, Republic of Korea
[6]Department of Radiology, Chonnam National University Hwasun Hospital and Chonnam National University Medical School, Hwasun, Republic of Korea
[7]Department of Radiology, Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea
[8]Department of Radiology, Seoul St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea
[9]Department of Radiology, Korea University Anam Hospital, Seoul, Republic of Korea
[10]Department of Radiology, Incheon Sejong Hospital, Incheon, Republic of Korea
[11]Department of Radiology, Seoul National University Hospital, Seoul, Republic of Korea
[12]Institute of Medical and Biological Engineering, Seoul National University Medical Research Center, Seoul, Republic of Korea

Objective: To evaluate the accuracy of large language models (LLMs) in extracting Coronary Artery Disease-Reporting and Data System (CAD-RADS) 2.0 components from coronary CT angiography (CCTA) reports, and assess the impact of prompting strategies.

Materials and Methods: In this multi-institutional study, we collected 319 synthetic, semi-structured CCTA reports from six institutions to protect patient privacy while maintaining clinical relevance. The dataset included 150 reports from a primary institution (100 for instruction development and 50 for internal testing) and 169 reports from five external institutions for external testing. Board-certified radiologists established reference standards following the CAD-RADS 2.0 guidelines for all three components: stenosis severity, plaque burden, and modifiers. Six LLMs (GPT-4, GPT-4o, Claude-3.5-Sonnet, o1-mini, Gemini-1.5-Pro, and DeepSeek-R1-Distill-Qwen-14B) were evaluated using an optimized instruction with prompting strategies, including zero-shot or few-shot with or without chain-of-thought (CoT) prompting. The accuracy was assessed and compared using McNemar's test.

Results: LLMs demonstrated robust accuracy across all CAD-RADS 2.0 components. Peak stenosis severity accuracies reached 0.980 (48/49, Claude-3.5-Sonnet and o1-mini) in internal testing and 0.946 (158/167, GPT-4o and o1-mini) in external testing. Plaque burden extraction showed exceptional accuracy, with multiple models achieving perfect accuracy (43/43) in internal testing and 0.993 (137/138, GPT-4o, and o1-mini) in external testing. Modifier detection demonstrated consistently high accuracy (≥0.990) across most models. One open-source model, DeepSeek-R1-Distill-Qwen-14B, showed a relatively low accuracy for stenosis severity: 0.898 (44/49, internal) and 0.820 (137/167, external). CoT prompting significantly enhanced

Received: March 13, 2025 **Revised:** July 12, 2025 **Accepted:** July 13, 2025
*These authors contributed equally to this work.
Corresponding author: Chang Min Park, MD, PhD, Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, 101 Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea
• E-mail: morphius@snu.ac.kr

the accuracy of several models, with GPT-4 showing the most substantial improvements: stenosis severity accuracy increased by 0.192 (P < 0.001) and plaque burden accuracy by 0.152 (P < 0.001) in external testing.
Conclusion: LLMs demonstrated high accuracy in automated extraction of CAD-RADS 2.0 components from semi-structured CCTA reports, particularly when used with CoT prompting.
Keywords: Coronary CT angiography; CAD-RADS 2.0; Information extraction; Large language model; Prompting strategy

INTRODUCTION

Coronary artery disease is one of the most prevalent and important cardiovascular diseases with significant morbidity and mortality [1]. Management depends on patient symptoms, risk assessment, and coronary artery involvement [2,3]. Coronary CT angiography (CCTA) enables detailed visualization of plaque characteristics, stenosis severity, and functional significance, thereby guiding treatment strategies [4]. Numerous studies have underscored the value of CCTA in both stable and acute chest pain diagnoses [5-7].

Although high-quality radiological examinations depend on multiple factors, consistent reporting is crucial for clinical decision making [8,9]. Structured impressions enhance interreader agreement and diagnostic clarity, as observed in other standardized reporting systems [10-12]. Coronary Artery Disease-Reporting and Data System 2.0 (CAD-RADS 2.0) provides a framework with three main components: stenosis severity, plaque burden, and modifiers [13-15]. This framework enables systematic documentation and facilitates clear communication among healthcare providers [16]. Despite these standardization benefits, the additional time and effort required for the implementation of CAD-RADS has limited its clinical adoption [17]. Furthermore, significant variability in radiology reporting remains a persistent challenge [18,19], where reporting discrepancies can lead to misclassification of stenosis severity or overlooking high-risk plaque (HRP) features, thereby impacting decisions and patient outcomes [20]. These challenges underscore the need for reliable automated approaches to support standardized CCTA reporting and reduce the burden of implementing CAD-RADS.

Large language models (LLMs) have demonstrated promising potential in medical text analysis, particularly in extracting structured information from clinical narratives and radiology reports [21-29]. The model performance is enhanced through different prompt engineering strategies, including instruction prompting (zero-shot prompting), few-shot prompting with example pairs, and chain-of-thought (CoT) prompting, which guide step-by-step reasoning [30-33].

CoT prompting has demonstrated particular effectiveness in complex medical reasoning tasks by enabling models to explicitly articulate their decision-making processes [34-36], which is crucial for clinical acceptance and error identification. This approach breaks down complex diagnostic criteria into manageable steps, thereby improving the accuracy and interpretability. Such explicit reasoning is especially valuable in standardized reporting systems such as CAD-RADS, where the precise application of multiple criteria and thresholds is essential for accurate classification.

Initial applications of LLMs to CAD-RADS extraction have shown promise, with a recent single-center study demonstrating 0.870 accuracy in stenosis severity extraction using GPT-4o on 100 CCTA reports [37] and another study evaluating the comprehension of LLMs through multiple-choice questions [38]. However, these studies were limited in scope, focusing primarily on stenosis severity without exploring the full range of CAD-RADS 2.0 components or the impact of different prompting techniques. This study aimed to address this knowledge gap by comprehensively evaluating the accuracy of LLMs in extracting all CAD-RADS 2.0 components from CCTA reports and assessing the impact of prompting strategies, particularly CoT prompting.

MATERIALS AND METHODS

This study received a waiver of approval after initial review by the Institutional Review Board of Seoul National University Boramae Medical Center, the primary study site (IRB No. 07-2024-32). As the study utilized entirely synthetic, non-identifiable data, additional IRB approvals from the other participating institutions were not required. Further details are provided in Supplementary Text 1. The code used in this study is available at https://github.com/reonaledo/cad-rads-extraction. The overall workflow of this study is shown in Figure 1.

Data

We collected synthetic CCTA reports from six cardiothoracic

Korean Journal of Radiology
KJR

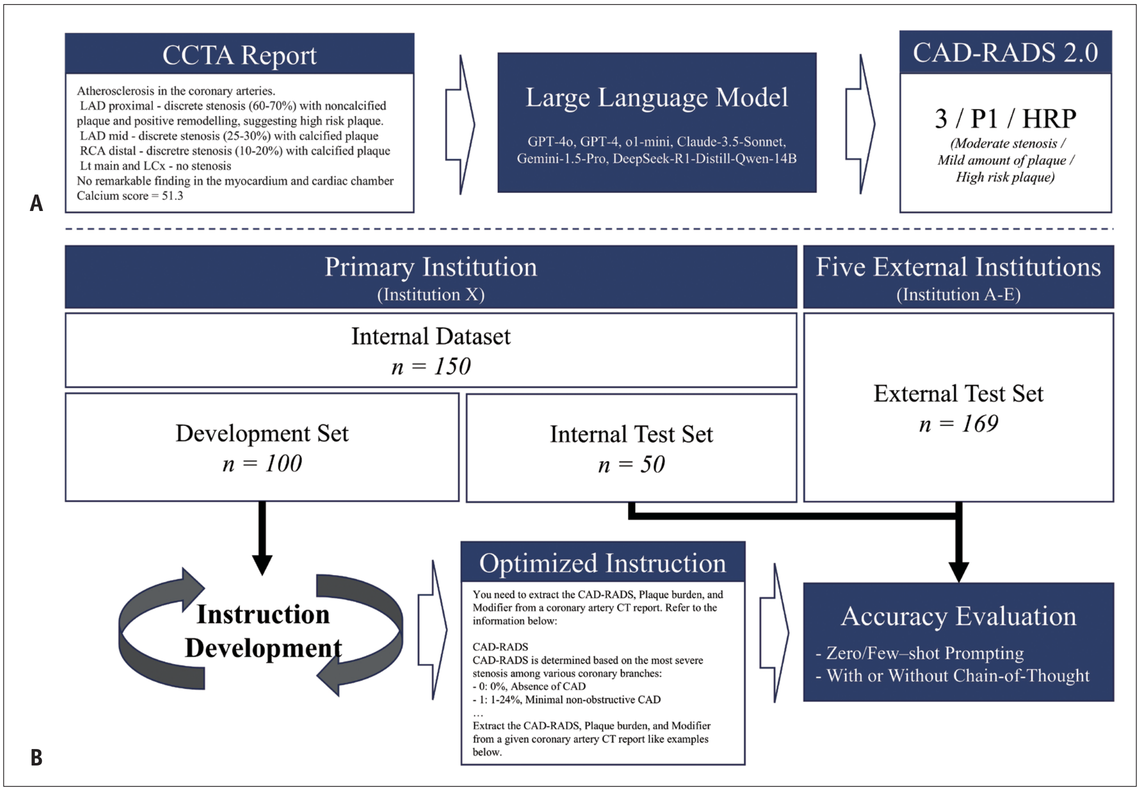

Fig. 1. Schematic flow diagram of the model evaluation process. **A:** Example of CCTA report, large language models for CAD-RADS 2.0 extraction, and example of CAD-RAD 2.0 including stenosis severity, plaque burden, and modifiers. **B:** Study datasets, example of instruction development, and accuracy evaluation process. CCTA = coronary CT angiography, CAD-RADS = Coronary Artery Disease-Reporting and Data System, HRP = high-risk plaque

radiologists (K.N.J., W.G.J., H.L., K.S.B., S.H.H., and E.Y.K., with 16, 6, 16, 10, 12, and 15 years of experience in cardiothoracic radiology, respectively) from different institutions in a single country. The dataset comprised 319 reports: 150 from the primary study site (institution X) and 169 from the remaining five external institutions (institutions A-E). The reports were mostly semi-structured, and representative examples are provided in Supplementary Table 1.

To protect patient privacy while maintaining clinical relevance, these synthetic reports were generated based on actual cardiac CT reports but were extensively modified by altering key clinical elements such as lesion location, extent, and plaque characteristics. Each radiologist created synthetic reports according to their institution's characteristic reporting style and format while preserving the natural variation in reporting patterns across different clinical settings. No actual clinical information, including patient sex, age, or symptoms, was utilized during generation, ensuring complete anonymization.

Of the 150 reports from institution X, 100 were designated as the development set for instruction development. The remaining 50 reports, along with the 169 reports from external institutions, were allocated to the internal and external test sets, respectively, for model evaluation.

Data Annotation

An overview of the data annotation process and establishment of a reference standard is shown in Figure 2. Two board-certified cardiothoracic radiologists (K.N.J. and M.Y.K.; 16 and 10 years of experience in radiology, respectively) reviewed all reports and established reference standards through consensus following CAD-RADS 2.0 guidelines. This process included the categorization of stenosis severity, plaque burden assessment, and evaluation of relevant modifiers.

Instruction Development

Two board-certified radiologists (C.M.P. and K.N.J.; 21 and 16 years of experience in radiology, respectively) selected three representative examples from the development set that encompassed the diversity of our dataset, and

 https://doi.org/10.3348/kjr.2025.0293

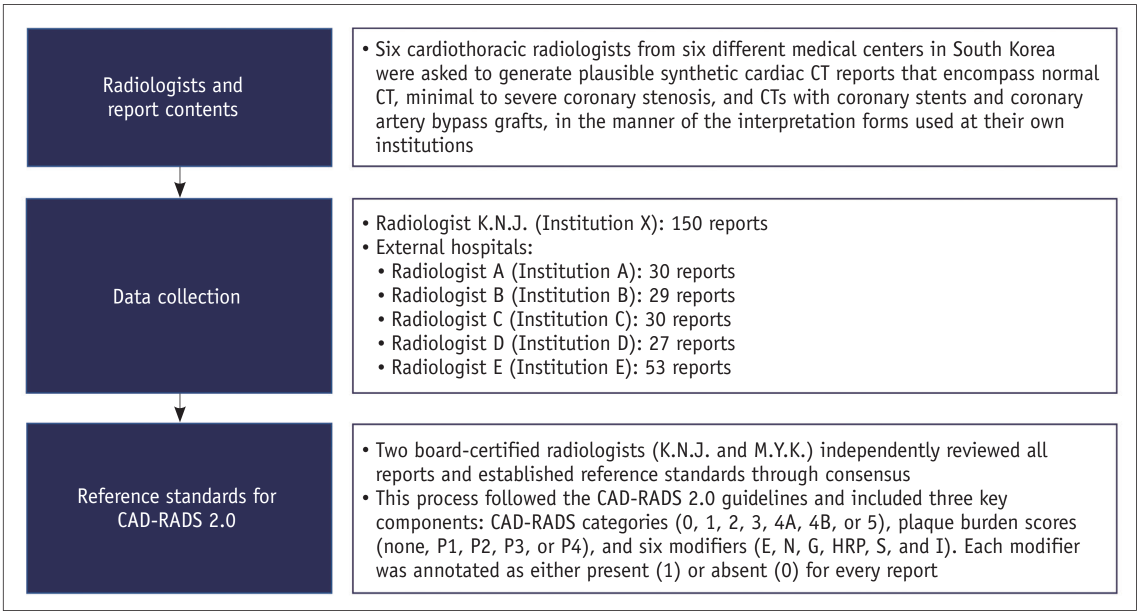

Fig. 2. Process of synthetic report generation and data collection for CCTA reports labeled with CAD-RADS 2.0. CCTA = coronary CT angiography, CAD-RADS = Coronary Artery Disease-Reporting and Data System, Institution X = Seoul National University Boramae Medical Center, Institution A = Incheon Sejong Hospital, Institution B = Korea University Anam Hospital, Institution C = Chonnam National University Hwasun Hospital, Institution D = Seoul St. Mary's Hospital, Institution E = Severance Hospital, E = exceptions, N = non-diagnostic study, G = grafts, HRP = high-risk plaque, S = stent, I = ischemia

provided input reports and their corresponding CAD-RADS components as reference cases for few-shot prompting. Based on these examples, they developed a comprehensive instruction incorporating the CAD-RADS 2.0 guidelines and definitions. The instruction was then refined through iterative testing using the development set to maximize the mean accuracy across all CAD-RADS components in the few-shot setting. For this process, we used three representative models (Claude-3.5-Sonnet, GPT-4o, and Gemini-1.5-Pro) and selected the instruction that achieved the highest average accuracy across these models. Once optimized, the instruction remained unchanged throughout all subsequent experiments (the detailed instruction structure is shown in Supplementary Fig. 1). The two radiologists had no prior exposure to the external test set (169 reports), which prevented data leakage.

Prompting Strategies

We evaluated the LLMs in two base prompt configurations using the optimized instruction: zero-shot prompting (without example cases) and few-shot prompting (with the three selected examples). To guide the model's step-by-step reasoning process in CAD-RADS 2.0 component extraction, we implemented CoT prompting by augmenting the base prompt with explicit reasoning paths. For few-shot CoT prompting, these three examples were enhanced with explicit reasoning paths, leading to CAD-RADS extraction, generated using Claude-3.5-Sonnet (example reasoning structure in Supplementary Fig. 2). For zero-shot CoT prompting, where example reasoning paths were not available, we followed the zero-shot CoT technique by adding "Let's think step by step." at the end of the prompt [33]. The detailed preparation process for the CoT example is described in Supplementary Text 2.

LLM Evaluation for CAD-RADS 2.0 Extraction

We evaluated five proprietary LLMs (GPT-4o, GPT-4, o1-mini, Claude-3.5-Sonnet, and Gemini-1.5-Pro) accessed through their official APIs in October 2024, and one open-source model (DeepSeek-R1-Distill-Qwen-14B; DeepSeek-14B) evaluated in a local environment. Each model received standardized inputs consisting of the optimized instruction, prompting strategy-specific prompts, and individual CCTA reports, and was required to output CAD-RADS 2.0 components (stenosis severity, plaque burden, and modifiers) in the JSON format. All evaluations followed the Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare guidelines [39], including setting the temperature to 0 and ensuring independence between CCTA reports. Further details of the experimental setup, including output formatting protocols, are provided in Supplementary Text 3.

The accuracy of the LLMs was assessed both overall and separately for three CAD-RADS 2.0 components: stenosis severity categories, plaque burden scores, and each of the six modifiers. Cases with modifier N (non-diagnostic) were excluded from the stenosis severity extraction accuracy, and cases with modifier S (stent) or without explicitly stated calcium scores were excluded from the plaque burden extraction accuracy.

Institution-Wise Analysis

To evaluate the generalizability of the model across different reporting patterns, we conducted detailed institution-wise accuracy analyses at five external institutions. We selected four representative LLMs: GPT-4o, which demonstrated the highest accuracy in our external test set; GPT-4, which has been extensively validated in various medical applications and is the most widely adopted model; Claude-3.5-Sonnet, which demonstrated the highest accuracy in our internal test set; and Gemini-1.5-Pro, which showed the lowest accuracy in our external test set.

Statistical Analysis

McNemar's test was used to assess the significance of accuracy differences between 1) pairwise model comparisons on identical test datasets and 2) models with and without CoT prompting. Statistical significance was set at $P < 0.05$. The Bonferroni correction was applied to account for multiple comparisons. All statistical analyses were performed by D.M. using Python (version 3.10.11; https://www.python.org) with statsmodels (version 0.14.2; https://www.statsmodels.org).

RESULTS

Dataset Characteristics

CCTA report generation and data collection are illustrated in Figure 2 and the distribution of CAD-RADS 2.0 components is shown in Table 1. Since this study used synthetic data, conventional demographic characteristics such as age and sex distribution could not be represented.

Table 1. Distribution of CAD-RADS 2.0 components across different datasets and institutions

CAD-RADS 2.0 components	Category	Internal dataset (n = 150)		External test dataset (n = 169)					
		Development set (n = 100)	Internal test set (n = 50)	Total (n = 169)	Institution A (n = 30)	Institution B (n = 29)	Institution C (n = 30)	Institution D (n = 27)	Institution E (n = 53)
Stenosis severity	0	9.0 (9)	6.0 (3)	5.9 (10)	6.7 (2)	20.7 (6)	0 (0)	7.4 (2)	0 (0)
	1	19.0 (19)	10.0 (5)	14.8 (25)	0 (0)	20.7 (6)	10.0 (3)	25.9 (7)	17.0 (9)
	2	21.0 (21)	38.0 (19)	26.0 (44)	43.3 (13)	20.7 (6)	20.0 (6)	18.5 (5)	26.4 (14)
	3	16.0 (16)	8.0 (4)	27.2 (46)	26.7 (8)	20.7 (6)	43.3 (13)	18.5 (5)	26.4 (14)
	4A	20.0 (20)	26.0 (13)	14.8 (25)	3.3 (1)	13.8 (4)	26.7 (8)	14.8 (4)	15.1 (8)
	4B	9.0 (9)	8.0 (4)	5.3 (9)	3.3 (1)	0 (0)	0 (0)	14.8 (4)	7.5 (4)
	5	6.0 (6)	4.0 (2)	5.9 (10)	16.7 (5)	3.4 (1)	0 (0)	0 (0)	7.5 (4)
Plaque burden	None	40.0 (40)	16.0 (8)	24.3 (41)	43.3 (13)	37.9 (11)	3.3 (1)	29.6 (8)	15.1 (8)
	P1	21.0 (21)	30.0 (15)	24.3 (41)	13.3 (4)	24.1 (7)	26.7 (8)	25.9 (7)	28.3 (15)
	P2	16.0 (16)	22.0 (11)	20.7 (35)	13.3 (4)	24.1 (7)	30.0 (9)	14.8 (4)	20.8 (11)
	P3	10.0 (11)	26.0 (13)	23.1 (39)	20.0 (6)	10.3 (3)	33.3 (10)	14.8 (4)	30.2 (16)
	P4	12.0 (12)	6.0 (3)	7.7 (13)	10.0 (3)	3.4 (1)	6.7 (2)	14.8 (4)	5.7 (3)
Modifiers	E	2.0 (2)	2.0 (1)	2.4 (4)	3.3 (1)	6.9 (2)	0 (0)	0 (0)	1.9 (1)
	N	14.0 (14)	2.0 (1)	1.2 (2)	0 (0)	0 (0)	0 (0)	0 (0)	3.8 (2)
	G	11.0 (11)	4.0 (2)	1.2 (2)	6.7 (2)	0 (0)	0 (0)	0 (0)	0 (0)
	HRP	0 (0)	10.0 (5)	4.1 (7)	0 (0)	3.4 (1)	6.7 (2)	7.4 (2)	3.8 (2)
	S	12.0 (12)	12.0 (6)	16.0 (27)	36.7 (11)	17.2 (5)	0 (0)	22.2 (6)	9.4 (5)
	I	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

Values are presented as percentage of total cases in each column with number of cases in parentheses. Age and sex demographics are not presented as this study utilized synthetic data generated from modified real-world data to protect patient privacy.
CAD-RADS = Coronary Artery Disease-Reporting and Data System, Institution A = Incheon Sejong Hospital, Institution B = Korea University Anam Hospital, Institution C = Chonnam National University Hwasun Hospital, Institution D = Seoul St. Mary's Hospital, Institution E = Severance Hospital, E = exceptions, N = non-diagnostic study, G = grafts, HRP = high-risk plaque, S = stent, I = ischemia

 https://doi.org/10.3348/kjr.2025.0293

In terms of stenosis severity, 2 and 4A categories were predominant in the internal test set, accounting for 64.0% of the cases. In the external test set, categories 0, 4B, and 5 each represented only 5.3%–5.9% of the cases. P4 plaque burden was relatively uncommon across both datasets, comprising 6.0% and 7.7% of the internal and external test sets, respectively. Notably, a HRP modifier was absent in the development set, while it appeared in 4.1%–10.0% of cases in the test sets, and no cases with the I modifier were present in any dataset.

LLM Accuracy in Extraction of CAD-RADS 2.0 Component: Overall

Table 2 presents the optimal accuracy achieved by each model across all evaluated prompting strategies, representing the clinical benchmark for automated CAD-RADS 2.0 extraction.

Peak Accuracy Summary

Proprietary LLMs showed robust accuracy across all components. The highest accuracies for stenosis severity were 0.980 (48/49, internal) and 0.946 (158/167, external). Plaque burden extraction showed exceptional results, with multiple models achieving perfect scores (43/43) in internal testing and up to 0.993 (137/138) in external testing. Modifier detection demonstrated a consistently high accuracy, with the mean accuracy across the six modifiers exceeding 0.989 in both test sets (49.5/50, internal; 167.2/169, external).

Analysis across all 36 model-component-test combinations revealed that few-shot CoT prompting achieved peak accuracy in 61.1% (22/36) of the cases, demonstrating its superiority over other approaches. When considering CoT strategies collectively (combining few-shot CoT and zero-shot CoT), they proved optimal in 86.1% (31/36) of the cases, confirming the substantial benefit of explicit reasoning guidance for CAD-RADS extraction tasks. Additionally, few-shot prompting methods were optimal in 75.0% (27/36) of the cases, further highlighting the value of providing examples to guide model performance.

Model-Specific Accuracy Characteristics

Claude-3.5-Sonnet and o1-mini achieved superior accuracy in internal testing, with Claude-3.5-Sonnet reaching 0.980 (48/49) stenosis accuracy, and o1-mini demonstrating consistent accuracy across datasets, achieving 0.980 (48/49) in internal testing and 0.946 (158/167) in external testing for stenosis severity. Open-source DeepSeek-14B demonstrated promising accuracy as an open-source alternative with stenosis accuracies of 0.898

Table 2. Peak accuracy summary across all prompting strategies

Model	Stenosis severity		Plaque burden		Modifiers (averaged)	
	Internal (n = 49)	External (n = 167)	Internal (n = 43)	External (n = 138)	Internal (n = 50)	External (n = 169)
GPT-4o	0.837 (41) (multiple methods)	0.940 (157) (multiple methods)	1.000 (43) (multiple methods)	0.993 (137) (few-shot CoT)	0.993 (49.7) (few-shot CoT)	1.000 (169.0) (few-shot CoT)
GPT-4	0.816 (40) (few-shot CoT)	0.928 (155) (few-shot CoT)	1.000 (43) (few-shot CoT)	0.971 (134) (few-shot CoT)	0.997 (49.8) (few-shot CoT)	0.989 (167.2) (few-shot CoT)
Claude-3.5-Sonnet	0.980 (48) (few-shot CoT)	0.904 (151) (few-shot)	1.000 (43) (few-shot CoT)	0.986 (136) (few-shot CoT)	0.993 (49.7) (few-shot CoT)	0.992 (167.7) (few-shot CoT)
Gemini-1.5-Pro	0.918 (45) (multiple methods)	0.856 (143) (few-shot CoT)	0.953 (41) (few-shot CoT)	0.942 (130) (zero-shot CoT)	0.997 (49.8) (zero-shot)	0.993 (167.8) (few-shot)
o1-mini	0.980 (48) (multiple methods)	0.946 (158) (multiple methods)	0.977 (42) (multiple methods)	0.993 (137) (few-shot CoT)	0.997 (49.8) (few-shot)	0.995 (168.2) (few-shot)
DeepSeek-14B	0.898 (44) (zero-shot CoT)	0.820 (137) (zero-shot CoT)	0.907 (39) (multiple methods)	0.935 (129) (zero-shot CoT)	0.990 (49.5) (few-shot CoT)	0.992 (167.7) (few-shot CoT)

Values are presented as accuracy with number of correct predictions in parentheses. For modifiers, values represent averaged accuracy across six modifiers with average number of correct predictions shown in parentheses. Parentheses also indicate the prompting method that achieved peak accuracy for each model-component combination. Multiple methods indicate tied accuracy. Internal test set: n = 49 for stenosis severity (excluding modifier N cases), n = 43 for plaque burden (excluding modifier S cases and cases without calcium scores), n = 50 for modifiers. External test set: n = 167 for stenosis severity (excluding modifier N cases), n = 138 for plaque burden (excluding modifier S cases and cases without calcium scores), n = 169 for modifiers. CoT = chain-of-thought, DeepSeek-14B = DeepSeek-R1-Distill-Qwen-14B

(44/49, internal) and 0.820 (137/167, external).

Statistical analysis using McNemar's test revealed significant accuracy differences between models under few-shot CoT prompting for stenosis severity and plaque burden but not for modifiers (*P* > 0.230). In the internal test set, Claude-3.5-Sonnet and o1-mini significantly outperformed GPT-4o in terms of stenosis severity (both *P* = 0.013). In external testing, GPT-4o significantly outperformed Gemini-1.5-Pro (*P* = 0.033) and DeepSeek-14B (*P* < 0.001) in terms of stenosis severity. When compared to DeepSeek-14B, all models except Gemini-1.5-Pro significantly outperformed DeepSeek-14B in terms of stenosis severity (all *P* < 0.012). For plaque burden assessment, GPT-4o, Claude-3.5-Sonnet, and o1-mini significantly outperformed DeepSeek-14B (all *P* < 0.007), whereas GPT-4 and Gemini-1.5-Pro showed no significant differences.

LLM Accuracy in Extraction of CAD-RADS 2.0 Component: Specific Components

Stenosis Severity

This was the most challenging component of the three tests, with accuracies ranging from 0.673 (33/49) to 0.980 (48/49) (Table 3). CoT effectiveness varied across models, with GPT-4 showing the most notable improvement through CoT, increasing the external accuracy by 0.126 (zero-shot, *P* < 0.001) and 0.192 (few-shot, *P* < 0.001). In contrast, o1-mini showed a minimal CoT response (difference = -0.041 to +0.020), suggesting pretrained reasoning capabilities. The accuracy consistency between the test sets varied substantially, with o1-mini maintaining stable accuracy (0.980 [48/49] to 0.946 [158/167]), whereas Gemini-1.5-Pro showed significant variation (0.918 [45/49] to 0.856 [143/167]).

Plaque Burden

Plaque burden extraction demonstrated superior accuracy compared to stenosis severity, with accuracies ranging from 0.698 (30/43) to 1.000 (138/138) (Table 4). CoT prompting showed substantial benefits, with Claude-3.5-Sonnet exhibiting notable improvement through zero-shot CoT (difference = +0.239, *P* < 0.001, external), and GPT-4 achieving perfect internal accuracy (43/43) with few-shot CoT (difference = +0.163, *P* = 0.016). Few-shot CoT prompting generally provided stable accuracy, with multiple proprietary models achieving perfect or near-perfect accuracy in both test sets: GPT-4o (1.000 [43/43], 1.000

Table 3. Stenosis severity extraction accuracy across prompting methods

Prompting method	GPT-4o		GPT-4		Claude-3.5-Sonnet		Gemini-1.5-Pro		o1-mini		DeepSeek-14B	
	Internal	External	Internal	External	Internal	External	Internal	External	Internal	External	Internal	External
Zero-shot	0.837 (41)	0.904 (151)	0.735 (36)	0.796 (133)	0.898 (44)	0.826 (138)	0.918 (45)	0.784 (131)	0.980 (48)	0.946 (158)	0.857 (42)	0.808 (135)
Zero-shot CoT	0.837 (41)	0.940 (157)	0.796 (39)	0.922 (154)	0.918 (45)	0.898 (150)	0.918 (45)	0.767 (128)	0.939 (46)	0.946 (158)	0.898 (44)	0.820 (137)
P-value	1.000	0.210	0.549	<0.001	1.000	0.023	1.000	0.743	0.625	1.000	0.625	0.831
Few-shot	0.816 (40)	0.880 (147)	0.673 (33)	0.737 (123)	0.898 (44)	0.904 (151)	0.714 (35)	0.773 (129)	0.959 (47)	0.922 (154)	0.837 (41)	0.707 (118)
Few-shot CoT	0.714 (35)	0.940 (157)	0.816 (40)	0.928 (155)	0.980 (48)	0.886 (148)	0.857 (42)	0.856 (143)	0.980 (48)	0.898 (150)	0.857 (42)	0.773 (129)
P-value	0.227	0.021	0.065	<0.001	0.125	0.664	0.065	0.020	1.000	0.481	1.000	0.035

Values are presented as accuracy with number of correct predictions in parentheses. Statistical significance was assessed using McNemar's test comparing each prompting method to its baseline (zero-shot for zero-shot CoT comparison, few-shot for few-shot CoT comparison). *P*-values <0.05 are considered statistically significant. Internal test set: n = 49 (excluding cases with modifier N). External test set: n = 167 (excluding cases with modifier N). CoT = chain-of-thought, DeepSeek-14B = DeepSeek-R1-Distill-Qwen-14B

Table 4. Plaque burden extraction accuracy across prompting methods

Prompting method	GPT-4o		GPT-4		Claude-3.5-Sonnet		Gemini-1.5-Pro		o1-mini		DeepSeek-14B	
	Internal	External	Internal	External	Internal	External	Internal	External	Internal	External	Internal	External
Zero-shot	0.977 (42)	0.971 (134)	0.930 (40)	0.870 (120)	0.884 (38)	0.703 (97)	0.698 (30)	0.783 (108)	0.953 (41)	0.978 (135)	0.907 (39)	0.855 (118)
Zero-shot CoT	1.000 (43)	0.993 (137)	0.907 (39)	0.862 (119)	0.884 (38)	0.942 (130)	0.814 (35)	0.942 (130)	0.977 (42)	0.986 (136)	0.907 (39)	0.935 (129)
P-value	1.000	0.250	1.000	1.000	1.000	<0.001	0.063	<0.001	1.000	1.000	1.000	0.019
Few-shot	1.000 (43)	0.993 (137)	0.837 (36)	0.819 (113)	0.953 (41)	0.971 (134)	0.907 (39)	0.855 (118)	0.977 (42)	0.978 (135)	0.884 (38)	0.732 (101)
Few-shot CoT	1.000 (43)	1.000 (138)	1.000 (43)	0.971 (134)	1.000 (43)	0.986 (136)	0.953 (41)	0.928 (128)	0.953 (41)	0.993 (137)	0.884 (38)	0.862 (119)
P-value	1.000	1.000	0.016	<0.001	0.500	0.500	0.688	0.041	1.000	0.500	1.000	<0.001

Values are presented as accuracy with number of correct predictions in parentheses. Statistical significance was assessed using McNemar's test comparing each prompting method to its baseline (zero-shot for zero-shot CoT comparison, few-shot for few-shot CoT comparison). *P*-values <0.05 are considered statistically significant. Internal test set: n = 43 (excluding cases with modifier S and cases without explicitly stated calcium scores). External test set: n = 138 (excluding cases with modifier S and cases without explicitly stated calcium scores). CoT = chain-of-thought, DeepSeek-14B = DeepSeek-R1-Distill-Qwen-14B

Table 5. Modifier extraction accuracy across prompting methods (averaged across all modifiers)

Prompting method	GPT-4o		GPT-4		Claude-3.5-Sonnet		Gemini-1.5-Pro		o1-mini		DeepSeek-14B	
	Internal	External	Internal	External	Internal	External	Internal	External	Internal	External	Internal	External
Zero-shot	0.973 (48.7)	0.985 (166.5)	0.940 (47.0)	0.948 (160.2)	0.973 (48.7)	0.972 (164.3)	0.997 (49.8)	0.992 (167.7)	0.993 (49.7)	0.992 (167.7)	0.977 (48.8)	0.975 (164.8)
Zero-shot CoT	0.980 (49.0)	0.984 (166.3)	0.977 (48.8)	0.976 (165.0)	0.987 (49.3)	0.975 (164.8)	0.983 (49.2)	0.982 (166.0)	0.990 (49.5)	0.994 (168.0)	0.983 (49.2)	0.979 (165.5)
P-value	0.375	1.000	0.289	<0.001	0.500	0.011	0.500	0.016	1.000	0.625	1.000	0.500
Few-shot	0.977 (48.9)	0.991 (167.5)	0.973 (48.7)	0.945 (159.7)	0.907 (45.3)	0.918 (155.2)	0.993 (49.7)	0.993 (167.8)	0.997 (49.8)	0.995 (168.2)	0.880 (44.0)	0.882 (149.0)
Few-shot CoT	0.993 (49.7)	1.000 (169.0)	0.997 (49.8)	0.989 (167.2)	0.993 (49.7)	0.992 (167.7)	0.990 (49.5)	0.986 (166.7)	0.993 (49.7)	0.994 (168.0)	0.990 (49.5)	0.992 (167.7)
P-value	0.500	0.031	0.125	<0.001	<0.001	<0.001	1.000	0.016	1.000	1.000	<0.001	<0.001

Values are presented as accuracy averaged across six modifiers (E, N, G, HRP, S, I) with average number of correct predictions shown in parentheses. Statistical significance was assessed using McNemar's test comparing each prompting method to its baseline (zero-shot for zero-shot CoT comparison, few-shot for few-shot CoT comparison); *P*-values represent the lowest *P*-value among the six modifiers. *P*-values <0.05 are considered statistically significant. Internal test set: n = 50. External test set: n = 169. E = exceptions, N = non-diagnostic study, G = grafts, HRP = high-risk plaque, S = stent, I = ischemia, CoT = chain-of-thought, DeepSeek-14B = DeepSeek-R1-Distill-Qwen-14B

[138/138]), GPT-4 (1.000 [43/43], 0.971 [134/138]), and Claude-3.5-Sonnet (1.000 [43/43], 0.986 [136/138]).

Modifiers

Modifiers achieved the highest accuracies among all the components, ranging from 0.880 (44.0/50) to 1.000 (169.0/169) (Table 5). Several models achieved near-perfect baseline accuracy with zero-shot prompting, indicating high intrinsic capability. Despite the high baseline accuracy, Claude-3.5-Sonnet and DeepSeek-14B showed significant CoT benefits, with DeepSeek-14B improving from 0.880 (44.0/50) to 0.990 (49.5/50, $P < 0.001$) during internal testing. Modifier extraction demonstrated the smallest accuracy gap between the internal and external test sets (less than 0.028), indicating its robust generalizability. Confusion matrices illustrating the classification patterns of the representative models under few-shot CoT prompting are shown in Supplementary Figures 3-5.

Institution-Wise Accuracy

As shown in Figure 3, institution-wise analysis using few-shot CoT prompting revealed a generally consistent accuracy across institutions B through E, with accuracies ranging from 0.815 (22/27) to 1.000 (30/30) for stenosis severity. Institution A showed a notably lower accuracy in stenosis severity extraction across all models, with Claude-3.5-Sonnet demonstrating an accuracy of 0.700 (21/30), and DeepSeek-14B showing an accuracy of 0.300 (9/30). For plaque burden, GPT-4o achieved accuracy of 1.000 (138/138) across all institutions, while GPT-4 and Claude-3.5-Sonnet maintained accuracies ≥0.956 (43/45). DeepSeek-14B showed variable accuracy, with perfect accuracy at Institutions A (18/18) and B (24/24) but reduced accuracy (0.767 [23/30] to 0.822 [37/45]). The decline in accuracy for stenosis severity at institution A is likely attributable to its distinct reporting style (Supplementary Table 1).

Error Analysis

Analysis of model errors across four representative LLMs (GPT-4, Claude-3.5-Sonnet, Gemini-1.5-Pro, and DeepSeek-14B) revealed five predominant error types in stenosis severity extraction: numerical threshold misapplication, qualitative term misinterpretation, multivessel classification errors, omission of explicit diagnostic criteria, and interpretational ambiguity (Table 6). The most frequent errors included numerical threshold misapplication (Gemini-1.5-Pro, 3.7%), qualitative term misinterpretation (DeepSeek-14B, 10.6%), and omission of explicit diagnostic criteria (GPT-4 and DeepSeek-14B, 3.2%). Representative examples of these error patterns are listed in Table 7.

DISCUSSION

We systematically evaluated the accuracy of various LLMs in extracting CAD-RADS 2.0 components from multicenter CCTA reports.

Our study is the first comprehensive evaluation of the accuracy of LLMs in accordance with CAD-RADS 2.0, which encompasses not only stenosis severity, but also plaque burden assessment and various modifiers. Through a systematic evaluation of multiple prompting strategies ranging from zero-shot to few-shot CoT approaches, proprietary LLMs demonstrated robust extraction accuracy across all CAD-RADS 2.0 components. Peak accuracy reached 0.980 (48/49) and 0.946 (158/167) for stenosis severity in the internal and external test sets, respectively, with multiple models achieving perfect accuracy (43/43) for plaque burden in internal testing and up to 0.993 (137/138) in external testing. The open-source model DeepSeek-14B showed promising accuracy as an open-source alternative with stenosis accuracies of 0.898 (44/49, internal) and 0.820 (137/167, external) and demonstrated robust modifier extraction capabilities (0.990 [49.5/50], internal; 0.992 [167.7/169], external), suggesting the potential for accessible, cost-effective solutions in clinical implementations.

Previous studies demonstrated the potential of LLMs in CAD-RADS extraction [37,38]. Silbergleit et al. [37] reported an accuracy of 0.870 for CAD-RADS scores using LLMs; however, their study was conducted using a relatively small dataset (n = 100) from a single institution and focused solely on stenosis severity without incorporating advanced prompt engineering techniques. Çamur et al. [38] evaluated LLMs through multiple-choice questions based on CAD-RADS 2.0 guidelines, but these questions were not derived from actual radiologic reports.

An analysis of the prompt engineering effectiveness provides practical insights. CoT prompting demonstrated variable effectiveness across different models and components. GPT-4 showed the most substantial improvements with CoT prompting in the few-shot setting, exhibiting accuracy increases for stenosis severity of 0.143 (internal) and 0.192 (external, $P < 0.001$) along

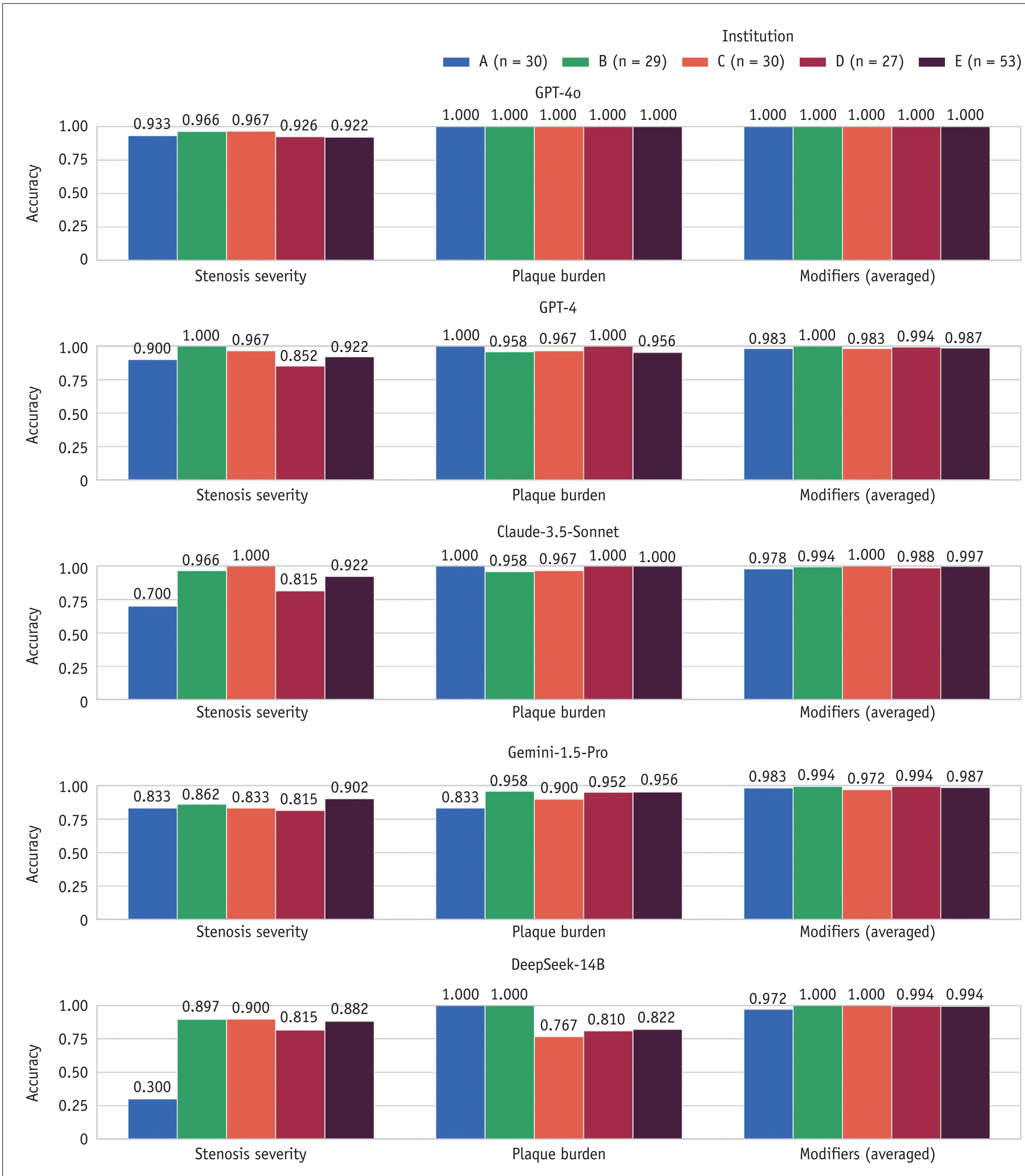

Fig. 3. Institution-specific analysis of model accuracy in CAD-RADS 2.0 component extraction across five external institutions. Evaluated models were GPT-4o, GPT-4, Claude-3.5-Sonnet, Gemini-1.5-Pro, and DeepSeek-14B. All results were obtained using few-shot chain-of-thought prompting. CAD-RADS = Coronary Artery Disease-Reporting and Data System, DeepSeek-14B = DeepSeek-R1-Distill-Qwen-14B, Institution A = Incheon Sejong Hospital, Institution B = Korea University Anam Hospital, Institution C = Chonnam National University Hwasun Hospital, Institution D = Seoul St. Mary's Hospital, Institution E = Severance Hospital

Table 6. Distribution of error types in extraction of stenosis severity

Error type	GPT-4 (errors = 21)	Claude-3.5-Sonnet (errors = 20)	Gemini-1.5-Pro (errors = 27)	DeepSeek-14B (errors = 45)
Numerical threshold misapplication	2.3 (5)	0.5 (1)	3.7 (8)	3.2 (7)
Qualitative term misinterpretation	0.9 (2)	4.2 (9)	2.3 (5)	10.6 (23)
Multi-vessel classification errors	1.4 (3)	0.9 (2)	0.5 (1)	0.9 (2)
Omission of explicit diagnostic criteria	3.2 (7)	1.9 (4)	2.3 (5)	3.2 (7)
Interpretational ambiguity	1.4 (3)	1.4 (3)	2.3 (5)	0.9 (2)
Others	0.5 (1)	0.5 (1)	1.4 (3)	1.9 (4)

Values are presented as percentage of total 216 test cases with number of errors in parentheses. The test cohort comprised 49 cases from internal test set and 167 cases from external test set, excluding cases with modifier N.
DeepSeek-14B = DeepSeek-R1-Distill-Qwen-14B

with notable plaque burden improvements of 0.163 (internal, $P = 0.016$) and 0.152 (external, $P < 0.001$). Zero-shot CoT also proved effective for GPT-4, improving the external stenosis accuracy by 0.126 ($P < 0.001$). Similarly, Claude-3.5-Sonnet benefited significantly from zero-shot CoT in plaque burden extraction, achieving a remarkable improvement of 0.239 in external testing ($P < 0.001$). In contrast, reasoning-enhanced models showed divergent patterns: o1-mini demonstrated minimal response to CoT prompting across all conditions (difference = -0.041 to +0.020), possibly due to its multi-step internal reasoning architecture, while DeepSeek-14B, which employs single-inference reasoning, showed substantial CoT benefits particularly in modifier extraction (improvement of 0.110 in both test sets, $P < 0.001$). These contrasting patterns suggest that the effectiveness of CoT prompting may depend on the underlying reasoning architecture, although further research is required to validate this hypothesis and develop tailor-prompting strategies for different model types.

The error analysis revealed systematic limitations in the application of CAD-RADS 2.0 by LLMs, even with few-shot CoT prompts. We identified five error types, of which three were the most prevalent: numerical threshold misapplication (3.7% in Gemini-1.5-Pro), qualitative term misinterpretation (10.6% in DeepSeek-14B), and omission of explicit diagnostic criteria (3.2% in GPT-4 and DeepSeek-14B). A detailed analysis of representative cases demonstrated these limitations. In qualitative term misinterpretation, models incorrectly categorized “moderate stenosis” as CAD-RADS 2 despite clear definitional guidelines specifying it as CAD-RADS 3. This error was particularly pronounced for DeepSeek-14B, potentially reflecting the impact of the model scale. In cases of diagnostic criteria omission, models often failed to apply CAD-RADS 4B classification when left main stenosis was reported as 50%–60%, despite explicit instructions specifying that left main stenosis ≥50% warrants a 4B classification. The models also exhibited numerical threshold errors, such as misclassification of a 69% stenosis into the 70%–99% range. These consistent patterns across models suggest that despite their promising capabilities, current LLMs still struggle with the reliable application of standardized diagnostic criteria, necessitating careful validation before clinical implementation.

The demonstrated performance of LLMs in CAD-RADS 2.0 extraction suggests a notable potential for clinical integration. When integrated into picture archiving and communications system (PACS) environments, these models automatically generate CAD-RADS values from radiology reports, enabling immediate verification and correction during report entry. The system can also cross-reference radiologist-provided assessments with model-generated values and flag discrepancies for quality assurance. Furthermore, with recent advancements in automated coronary analysis software that generate structured reports from CCTA images [40], our approach can leverage these outputs to produce CAD-RADS classifications directly from automated reports. This would enable the automatic display of standardized CAD-RADS information on PACS worklists immediately upon scan acquisition, even before formal radiologist review. This seamless integration could enhance efficiency, support immediate patient care prioritization, and improve the reliability of CAD-RADS reporting.

Our study has several limitations. First, although our CCTA reports were generated by experienced cardiothoracic radiologists to reflect daily clinical practice scenarios, they may not fully capture the complexity and variability of real-world clinical reports. Second, although our multi-institutional dataset provides insights into the model’s generalizability, the relatively small sample size (219 reports) and single-country origin may limit the broader

Table 7. Representative examples of systematic error patterns in extraction of stenosis severity

No	Error type	Model	Institution	Report	GT	Model's rationale (chain-of-thought)
1	Numerical threshold misapplication	Gemini-1.5-Pro	Severance Hospital	… RCA: Calcified plaques in the p and mRCA with moderate stenosis (60%–70%). …	3	Let's think step by step. 1. CAD-RADS: - The most severe stenosis mentioned is "moderate stenosis (60-70%)" in the RCA. - When presented with a range like '60%–70%', we interpret it as the upper end of the range minus 1, 69%. - This falls into the CAD-RADS 4A category: 70%–99% in a single or 2-vessel, Severe stenosis. …
2	Qualitative term misinterpretation	Claude-3.5-Sonnet	Chonnam National University Hwasun Hospital	… 1. p- to m-LAD: tubular noncalcified plaque, with moderate stenosis and positive remodeling (high-risk plaque). p-RCA: discrete partially calcified plaque, with moderate stenosis. …	3	Let's think step by step. 1. CAD-RADS: - The report mentions "p- to m-LAD: tubular noncalcified plaque, with moderate stenosis and positive remodeling (high-risk plaque)." - "Moderate stenosis" falls into CAD-RADS 2: 25%–49%, Mild non-obstructive CAD. …
3	Multi-vessel classification errors	GPT-4	Seoul National University Boramae Medical Center	… LM, pLAD, unremarkable. mLAD, pLCx, dOM, pRCA 70%–80% stenosis with mixed plaque. dLAD, mRCA 40%–50% stenosis with mixed plaque. …	4B	let's think step by step. 1. CAD-RADS: - The report mentions "mLAD, pLCx, dOM, pRCA 70%–80% stenosis with mixed plaque." - When presented with a range like '70%–80%', we interpret it as the upper end of the range minus 1, so 79%. - This falls into the CAD-RADS 4A category: 70%–99% in a single or 2-vessel, Severe stenosis. …
4	Omission of explicit diagnostic criteria	DeepSeek-14B	Severance Hospital	… LM: Calcified plaques with moderate stenosis. (50%–60%) LAD: Calcified plaques in the pLAD and DG with severe stenosis. (70%–75%) Calcified plaques in the m and dLAD with moderate stenosis (60%–69%) LCx: No evidence of plaque or stenosis. …	4B	… First, looking at the CAD-RADS. The report mentions LM with moderate stenosis (50%–60%), LAD with severe stenosis (70%–75%), and RCA also with severe stenosis (70%–75%). According to the CAD-RADS criteria, 70%–75% falls into the 4A category because it's a single or two-vessel severe stenosis. However, I need to check if it meets 4B. The LM has 50%–60%, which is moderate, not severe. The LAD and RCA each have severe stenosis, but since they are two different vessels, it's still 4A. So CAD-RADS is 4A. …
5	Interpretational ambiguity	Gemini-1.5-Pro	Incheon Sejong Hospital	… 1. mid RCA (near AM branching portion): focal noncalcified plaque with moderate to severe stenosis. 2. LM to proximal LAD ; noncalcified plaque with moderate to severe stenosis S/P stent for proximal LAD ; probable severe in-stent restenosis or occlusion. …	4B	Let's think step by step. 1. CAD-RADS: - The report mentions "probable severe in-stent restenosis or occlusion" in the proximal LAD. - Occlusion implies 100% stenosis, corresponding to CAD-RADS 5. - Therefore, despite the other findings, the presence of probable occlusion leads to a CAD-RADS score of 5.

Examples demonstrating characteristic error patterns observed across different models. Each case represents a distinct type of reasoning failure: Case 1 shows incorrect application of numerical thresholds despite explicit percentage ranges; Case 2 demonstrates misinterpretation of qualitative descriptors, where "moderate stenosis" was classified as CAD-RADS 2 despite being explicitly defined as CAD-RADS 3 in the instructions; Case 3 illustrates failure to identify three-vessel disease pattern (≥70% stenosis in three vessels) despite clear evidence; Case 4 shows omission of the CAD-RADS 4B criterion for LM stenosis ≥50%; and Case 5 represents a case where ambiguous report phrasing ("probable severe in-stent restenosis or occlusion") led to model interpreting it as total occlusion (CAD-RADS 5) while ground truth was determined as CAD-RADS 4B.
CAD-RADS = Coronary Artery Disease-Reporting and Data System, LM = left main, GT = Ground Truth, RCA = right coronary artery, LAD = left anterior descending, LCx = left circumflex, OM = obtuse marginal, DG = diagonal branch, AM = acute marginal branch, S/P = status post, p- = proximal, m- = mid, d- = distal, DeepSeek-14B = DeepSeek-R1-Distill-Qwen-14B

applicability of our findings. Third, we could not evaluate the “ischemia” modifier (I) because the institutions did not perform CT-derived fractional flow reserve assessment, and plaque burden assessment was limited to cases with documented calcium scores. Fourth, as the reports from each institution were generated by a single cardiothoracic radiologist, our findings may not fully represent institutional reporting variability. Fifth, because we used external APIs provided by proprietary LLMs, our methodology may not be directly applicable in clinical practice because of data privacy and security concerns related to externally transmitting patient data [41]. Finally, given the rapid pace of technological advancements, the models evaluated in this study have already been succeeded by newer versions (e.g., Gemini 2.5), potentially limiting the generalizability and long-term applicability of our results.

In conclusion, our multi-institutional study demonstrated that LLMs can achieve high accuracy in extracting CAD-RADS 2.0 components from semi-structured CCTA reports, particularly when used with CoT-prompting techniques. These findings suggest significant potential for improving standardization and consistency in CCTA reporting, which could reduce the workflow burden and enhance diagnostic clarity.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2025.0293.

Availability of Data and Material

Data generated or analyzed during the study are available from the corresponding author by request.

Conflicts of Interest

Hye-Jeong Lee, a Section Editor of the *Korean Journal of Radiology*, was not involved in the editorial evaluation or decision to publish this article.

Kwang Nam Jin has received research grants from Lunit unrelated to the present article. Chang Min Park has received research grants from Lunit, Coreline Soft, and Monitor Corp, and holds stock in Lunit, Coreline Soft, Promedius, and stock options in Lunit, Coreline Soft, and Monitor Corp, all unrelated to the present article.

The remaining authors have declared no conflicts of interest.

Author Contributions

Conceptualization: Chang Min Park, Kwang Nam Jin. Data curation: Kwang Nam Jin, Dabin Min, SangHeum Bang, Moon Young Kim, Won Gi Jeong, Hye-Jeong Lee, Kyongmin Sarah Beck, Sung Ho Hwang, Eun Young Kim. Formal analysis: Dabin Min, Kwang Nam Jin. Funding acquisition: Chang Min Park. Investigation: Dabin Min, Kwang Nam Jin. Methodology: Dabin Min, Kwang Nam Jin. Project administration: Chang Min Park. Resources: Chang Min Park, Kwang Nam Jin. Software: Dabin Min. Supervision: Chang Min Park. Validation: Dabin Min, Kwang Nam Jin, SangHeum Bang, Moon Young Kim. Visualization: Dabin Min. Writing—original draft: Dabin Min, Kwang Nam Jin. Writing—review & editing: Chang Min Park, Kwang Nam Jin, Hack-Lyoung Kim, Kyongmin Sarah Beck.

ORCID IDs

Dabin Min
https://orcid.org/0009-0007-4207-7646
Kwang Nam Jin
https://orcid.org/0000-0001-5494-9113
SangHeum Bang
https://orcid.org/0009-0003-2239-0986
Moon Young Kim
https://orcid.org/0000-0003-3025-0409
Hack-Lyoung Kim
https://orcid.org/0000-0002-6703-1472
Won Gi Jeong
https://orcid.org/0000-0003-2821-2788
Hye-Jeong Lee
https://orcid.org/0000-0003-4349-9174
Kyongmin Sarah Beck
https://orcid.org/0000-0002-9262-1001
Sung Ho Hwang
https://orcid.org/0000-0003-1850-0751
Eun Young Kim
https://orcid.org/0000-0002-2101-7982
Chang Min Park
https://orcid.org/0000-0003-1884-3738

Funding Statement

This study was supported by the National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (MSIT) (Grant No. RS-2024-00354666) and supported by the Seoul National University Hospital Research Fund (Grant No. 03-2023-0410).

Acknowledgments

Claude-3.5-Sonnet was utilized for drafting and grammatical revision to achieve natural English expressions in the completion of the sentences.

REFERENCES

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study. *J Am Coll Cardiol* 2020;76:2982-3021
2. Virani SS, Newby LK, Arnold SV, Bittner V, Brewer LC, Demeter SH, et al. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA guideline for the management of patients with chronic coronary disease: a report of the American Heart Association/American College of Cardiology Joint Committee on clinical practice guidelines. *Circulation* 2023;148:e9-e119
3. Gulati M, Levy PD, Mukherjee D, Amsterdam E, Bhatt DL, Birtcher KK, 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:e368-e454
4. Blankstein R, Shaw LJ, Gulati M, Atalay MK, Bax J, Calnon DA, et al. Implications of the 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR chest pain guideline for cardiovascular imaging: a multisociety viewpoint. *JACC Cardiovasc Imaging* 2022;15:912-926
5. Bittner DO, Ferencik M, Douglas PS, Hoffmann U. Coronary CT angiography as a diagnostic and prognostic tool: perspective from a multicenter randomized controlled trial: PROMISE. *Curr Cardiol Rep* 2016;18:40
6. SCOT-HEART investigators. CT coronary angiography in patients with suspected angina due to coronary heart disease (SCOT-HEART): an open-label, parallel-group, multicentre trial. *Lancet* 2015;385:2383-2391
7. Litt HI, Gatsonis C, Snyder B, Singh H, Miller CD, Entrikin DW, et al. CT angiography for safe discharge of patients with possible acute coronary syndromes. *N Engl J Med* 2012;366:1393-1403
8. Meinel FG, Bayer RR 2nd, Zwerner PL, De Cecco CN, Schoepf UJ, Bamberg F. Coronary computed tomographic angiography in clinical practice: state of the art. *Radiol Clin North Am* 2015;53:287-296
9. Ghoshhajra BB, Lee AM, Ferencik M, Elmariah S, Margey RJ, Onuma O, et al. Interpreting the interpretations: the use of structured reporting improves referring clinicians' comprehension of coronary CT angiography reports. *J Am Coll Radiol* 2013;10:432-438
10. Tessler FN, Middleton WD, Grant EG, Hoang JK, Berland LL, Teefey SA, et al. ACR thyroid imaging, reporting and data system (TI-RADS): white paper of the ACR TI-RADS committee. *J Am Coll Radiol* 2017;14:587-595
11. Eghtedari M, Chong A, Rakow-Penner R, Ojeda-Fournier H. Current status and future of BI-RADS in multimodality imaging, from the AJR special series on radiology reporting and data systems. *AJR Am J Roentgenol* 2021;216:860-873
12. Christensen J, Prosper AE, Wu CC, Chung J, Lee E, Elicker B, et al. ACR lung-RADS v2022: assessment categories and management recommendations. *J Am Coll Radiol* 2024;21:473-488
13. Cury RC, Abbara S, Achenbach S, Agatston A, Berman DS, Budoff MJ, et al. CAD-RADS™: 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 Am Coll Radiol* 2016;13(12 Pt A):1458-1466.e9
14. Canan A, Ranganath P, Goerne H, Abbara S, Landeras L, Rajiah P. CAD-RADS: pushing the limits. *Radiographics* 2020;40:629-652
15. Cury RC, Leipsic J, Abbara S, Achenbach S, Berman D, Bittencourt M, et al. CAD-RADS™ 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 Cardiovasc Comput Tomogr* 2022;16:536-557
16. Ahmadzadeh K, Roshdi Dizaji S, Kiah M, Rashid M, Miri R, Yousefifard M. The value of coronary artery disease - reporting and data system (CAD-RADS) in outcome prediction of CAD patients; a systematic review and meta-analysis. *Arch Acad Emerg Med* 2023;11:e45
17. Dewey M. Structure or entropy in reporting cardiac CT findings. *Int J Cardiovasc Imaging* 2016;32:1657-1658
18. White T, Aronson MD, Sternberg SB, Shafiq U, Berkowitz SJ, Benneyan J, et al. Analysis of radiology report recommendation characteristics and rate of recommended action performance. *JAMA Netw Open* 2022;5:e2222549
19. Liu Y, Feng Z, Qin S, Yang J, Han C, Wang X. Structured reports of pelvic magnetic resonance imaging in primary endometrial cancer: potential benefits for clinical decision-making. *PLoS One* 2019;14:e0213928
20. Maroules CD, Hamilton-Craig C, Branch K, Lee J, Cury RC, Maurovich-Horvat P, et al. Coronary artery disease reporting and data system (CAD-RADSTM): inter-observer agreement for assessment categories and modifiers. *J Cardiovasc Comput Tomogr* 2018;12:125-130
21. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023;620:172-180
22. OpenAI; Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. GPT-4 technical report. arXiv [Preprint]. 2023 [accessed on March 10, 2025]. Available at: https://doi.org/10.48550/arXiv.2303.08774
23. Anthropic. Claude 3.5 Sonnet [accessed on October 16, 2024]. Available at: https://www.anthropic.com/news/claude-3-5-sonnet

24. Gemini Team Google, Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. arXiv [Preprint]. 2024 [accessed on March 10, 2025]. Available at: https://doi.org/10.48550/arXiv.2403.05530
25. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv [Preprint]. 2025 [accessed on March 10, 2025]. Available at: https://doi.org/10.48550/arXiv.2501.12948
26. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* 2023;307:e230725
27. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of using the privacy-preserving large language model Vicuna for labeling radiology reports. *Radiology* 2023;309:e231147
28. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan R, et al. Privacy-preserving large language models for structured medical information retrieval. *NPJ Digit Med* 2024;7:257
29. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the diagnosis please quizzes. *Radiology* 2023;308:e231040
30. Ge Y, Guo Y, Das S, Al-Garadi MA, Sarker A. Few-shot learning for medical text: a review of advances, trends, and opportunities. *J Biomed Inform* 2023;144:104458
31. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. arXiv [Preprint]. 2022 [accessed on March 10, 2025]. Available at: https://doi.org/10.48550/arXiv.2201.11903
32. Liévin V, Hother CE, Motzfeldt AG, Winther O. Can large language models reason about medical questions? *Patterns (N Y)* 2024;5:100943
33. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv [Preprint]. 2022 [accessed on March 10, 2025]. Available at: https://doi.org/10.48550/arXiv.2205.11916
34. Miao J, Thongprayoon C, Suppadungsuk S, Krisanapan P, Radhakrishnan Y, Cheungpasitporn W. Chain of thought utilization in large language models and application in nephrology. *Medicina (Kaunas)* 2024;60:148
35. Wang Z, Zhang Z, Traverso A, Dekker A, Qian L, Sun P. Assessing the role of GPT-4 in thyroid ultrasound diagnosis and treatment recommendations: enhancing interpretability with a chain of thought approach. *Quant Imaging Med Surg* 2024;14:1602-1615
36. Ting YT, Hsieh TC, Wang YF, Kuo YC, Chen YJ, Chan PK, *et al.* Performance of ChatGPT incorporated chain-of-thought method in bilingual nuclear medicine physician board examinations. *Digit Health* 2024;10:20552076231224074
37. Silbergleit M, Tóth A, Chamberlin JH, Hamouda M, Baruah D, Derrick S, et al. ChatGPT vs Gemini: comparative accuracy and efficiency in CAD-RADS score assignment from radiology reports. *J Imaging Inform Med* 2024 Nov 11 [Epub]. Available at: https://doi.org/10.1007/s10278-024-01328-y
38. Çamur E, Cesur T, Güneş YC. Can large language models be new supportive tools in coronary computed tomography angiography reporting? *Clin Imaging* 2024;114:110271
39. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol* 2024;25:865-868
40. Kay FU, Canan A, Kukkar V, Hulsey K, Scanio A, Fan C, et al. Diagnostic accuracy of on-premise automated coronary CT angiography analysis based on coronary artery disease reporting and data system 2.0. *Radiology* 2025;315:e242087
41. Ng KKY, Matsuba I, Zhang PC. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. *NEJM AI* 2025;2:AIra2400380