Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Diagnostic Performance of Expert Physicians Versus General-Purpose Artificial Intelligence Using Standardized Static Coronary CT Images: A Dual-Reference Validation Study

Sefa Okar
drsefaokar@gmail.com
Başkent University Hospital
Ziya Gökalp BİLGEL
Başkent University Hospital
İsa Göktürk BALCI
Başkent University Hospital
Gürcan ERBAY
Başkent University Hospital
Mustafa YILMAZ
Başkent University Hospital

Research Article

Keywords: Coronary CT Angiography (CCTA), Artificial Intelligence, Large Language Models (LLM), Quantitative Coronary Angiography (QCA), Diagnostic Accuracy, Machine Learning

Posted Date: April 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9135888/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background: Coronary CT angiography (CCTA) is a first-line diagnostic modality for coronary artery disease (CAD), yet its interpretation requires significant expert experience. Although general-purpose multimodal artificial intelligence (GP-AI) models have shown promise in text-based medical tasks, their visual diagnostic performance in evaluating complex CCTA data remains poorly defined.

Methods: This single-center retrospective study included 63 patients (252 vessel-based image sets) who underwent both CCTA and invasive coronary angiography. Expert physician consensus and four frontier GP-AI models (GPT-4o, Gemini 2.5, Claude 3.5 Sonnet, and Grok 4) evaluated identical standardized static images using a zero-shot approach with default generation parameters. Obstructive disease was defined as ≥ 50% luminal stenosis. Diagnostic performance was validated against expert consensus for plaque characterization and quantitative coronary angiography (QCA) for stenosis severity.

Results: Expert consensus demonstrated robust agreement with QCA across all coronary territories (κ = 0.774–0.933, $p < 0.001$). In contrast, a marked performance disparity was observed for the GP-AI models; none achieved statistically significant agreement with QCA in the prognostically critical left anterior descending (LAD) or left main coronary arteries (LMCA) ($p > 0.05$). While Gemini 2.5 showed a moderate correlation in the right coronary artery (ICC = 0.515), overall continuous stenosis assessment and plaque characterization remained uniformly limited and clinically unreliable across all models.

Conclusion: Expert physician interpretation remains the reference standard for CCTA. Current frontier GP-AI models are not suitable for independent clinical interpretation of coronary imaging, particularly in anatomically complex segments. These findings emphasize that general visual reasoning cannot yet replace domain-specific cardiovascular AI solutions or expert clinical judgment in specialized radiological tasks.

1. INTRODUCTION

Coronary artery disease (CAD) remains the leading cause of morbidity and mortality globally [1]. Due to its ability to provide non-invasive imaging of coronary atherosclerosis and stenosis, recent guidelines have established Coronary CT Angiography (CCTA) as the "first-line test" [1–3] and the "cornerstone" of diagnostic management [4]. Beyond merely identifying lumen stenosis, CCTA provides critical data regarding coronary anatomy and plaque characterization [4, 5]. Currently, the interpretation of these images relies heavily on the expertise of specialized cardiologists and radiologists [6]. Recent reviews highlight the growing potential of AI-based clinical decision support systems in optimizing cardiovascular disease management and risk assessment [7]. However, as reported in previous studies, visual assessment of CCTA is time-consuming, highly dependent on reader experience, and hallmarked by significant interobserver variability [4, 6, 8–10].

To address these limitations, artificial intelligence (AI) technologies are increasingly being integrated into cardiovascular imaging [6]. "Narrow" AI tools specifically trained for CCTA—such as Cleerly and HeartFlow—have successfully entered clinical workflows, offering automated vessel segmentation and stenosis grading [4, 11, 12]. Extensive clinical trials have demonstrated that these purpose-built platforms enhance diagnostic consistency and reduce reading times.

In contrast, the role of general-purpose multimodal artificial intelligence (GP-AI) like ChatGPT, Gemini, Grok, and Claude in medical imaging remains a subject of intense debate. The vast majority of existing studies have evaluated these models solely on text-based tasks [13]. The only notable study testing visual diagnostic capability was limited to pediatric chest radiographs, where models performed at a "chance level," failing to demonstrate consistent radiological reasoning [14].

To date, there is virtually no data rigorously evaluating the ability of these multimodal models to interpret coronary angiography. This study aims to bridge that specific gap. Moving beyond text-based benchmarking,

we directly tested the visual diagnostic performance of four frontier GP-AI models (Gemini 2.5, Grok 4, Claude 3.5 Sonnet, and ChatGPT-4o). In our study design, both the AI models and the expert physician consensus were presented with the exact same standardized static images (curved multiplanar reconstruction (cMPR)). We then evaluated the diagnostic accuracy of their interpretations against two rigorous reference standards: the full clinical CCTA reports and invasive quantitative coronary angiography (QCA).

2. MATERIALS AND METHODS

2.1. Study Design and Study Population This study was designed as a retrospective diagnostic accuracy investigation conducted at a single tertiary referral center between May 2022 and December 2024. Consecutive patients who underwent Coronary Computed Tomography Angiography (CCTA) for clinical indications and subsequently received invasive coronary angiography (ICA) within 30 days were screened for inclusion (n = 66). To ensure a methodologically rigorous evaluation focused exclusively on native coronary artery anatomy, patients with a history of prior coronary stent implantation or coronary artery bypass grafting (CABG) were excluded (n = 3). Following the application of these exclusion criteria, a total of 63 patients constituted the final study cohort. All procedures were performed in accordance with the principles of the Declaration of Helsinki. This study was approved by Baskent University Institutional Review Board (Project no: KA25/407) and supported by Baskent University Research Fund. Given the retrospective nature of the study, the requirement for written informed consent was waived by the institutional review board.

2.2. Imaging Protocol and Coronary Computed Tomography Angiography Acquisition. To create a standardized visual dataset for a fair comparison between human experts and GP-AI models, representative 2D cMPR keyframes were selected by a dedicated 'preparation team' of two cardiovascular imaging specialists. This team was unblinded, with full access to patients' complete 3D CCTA volumetric datasets, clinical histories, original radiology reports, and reference invasive coronary angiography (ICA) results. Their objective was to deliberately isolate the optimal static frame demonstrating the most severe lesion or characteristic pathology for each vessel.

Subsequently, these optimized static images were evaluated by two independent, blinded expert readers. Having no access to the clinical context, 3D CCTA volumes, or ICA results, these physicians reviewed the images under strictly identical, constrained visual conditions as the GP-AI models. This two-stage design effectively eliminated clinical and spatial context as confounding variables

2.3. Image Standardization and Selection of Representative Keyframes

To enable a fair and objective comparison between human experts and general-purpose artificial intelligence (GP-AI) models, a 'Representative Keyframe Selection Protocol' was implemented to create a standardized visual dataset (Fig. 1). In the preparatory phase of this process, a dedicated preparation team consisting of a Level-3 cardiovascular radiologist and a Level-2 cardiologist reviewed the scans, providing full (unblinded) access to all 3D CCTA volumetric datasets, clinical histories, original radiology reports, and reference invasive coronary angiography (ICA) results of the patients.

The primary goal of this team was to intentionally isolate two high-resolution static cMPR frames that optimally showed the most severe lesion or characteristic pathology for each major epicardial coronary artery (LAD, LCx, and RCA). These selected images (JPEG) were completely anonymized by removing all patient identifiers and radiological markings.

The main diagnostic evaluation phase of the study was performed in a blinded manner by two independent expert readers (a radiologist and a cardiologist). These physicians, having no access to the patients' clinical context, 3D CCTA volumes, or reference ICA results, evaluated the images under completely identical and

restricted visual conditions, just like the GP-AI models. This two-stage design successfully eliminated the influence of clinical and spatial context as confounding variables and ensured that the pure visual recognition capabilities of both human experts and GP-AI models were tested against a standard reference. In total, 252 vessel-specific image sets (four vessels per patient across 63 patients) were prepared and subjected to evaluation by both GP-AI models and expert readers.

2.4. GP-AI Models and Evaluation Methodology

In this study, four frontier general-purpose multimodal artificial intelligence (GP-AI) models were evaluated using their most advanced versions available as of February 2026: GPT-4o (OpenAI), Gemini 2.5 (Google), Claude 3.5 Sonnet (Anthropic), and Grok 4 (xAI) [15–17]. All models were accessed through their official web-based interfaces using default generation parameters. To ensure methodological consistency, minimize stochastic variability, and address reproducibility concerns, each image was processed in a new and independent chat session with only a single iteration to prevent cross-contamination of data (zero-shot evaluation).

For each major epicardial coronary artery, both GP-AI models and blinded expert readers were required to report predefined diagnostic parameters based on the standardized image sets (Fig. 2). These parameters included: (i) the presence or absence of obstructive coronary stenosis (defined as $\geq 50\%$ luminal narrowing relative to reference vessel diameter), (ii) the estimated percentage of luminal narrowing, and (iii) the morphological classification of detected atherosclerotic plaques as calcified, non-calcified (soft), or mixed.

2.5. Reference Standards

To avoid subjective interpretation in the assessment of coronary artery lumen stenosis, Quantitative Coronary Angiography (QCA) was established as the reference standard. Invasive coronary angiography procedures were performed using the Siemens Artis zee angiography system. All angiographic images were analyzed by an independent interventional cardiologist blinded to CCTA and AI findings using syngo QCA software (Siemens Healthineers, Erlangen, Germany). In accordance with standard validation protocols [18], an automated edge-detection algorithm was used to determine arterial contours. Calibration was achieved using catheter tip contrast filling as a reference scaling tool. Minimal lumen diameter (MLD) and reference vessel diameter (RVD) were measured from the end-diastolic square, indicating the most severe stenosis and the least foreshortening (Fig. 3). A stenosis threshold of $\geq 50\%$ relative to the reference vessel diameter was defined as hemodynamically significant obstructive coronary artery disease. For plaque characterization, considering the limitations of conventional angiography in differentiating plaque components, expert physician consensus, validated by an experienced cardiologist and radiologist with full access to the patients' volumetric CCTA datasets, was accepted as the reference standard.

2.6. Statistical Analysis The diagnostic performance of artificial intelligence models and human readers was evaluated by calculating sensitivity, specificity, positive predictive value, and negative predictive value. Inter-method agreement for categorical variables was assessed using Cohen’s Kappa (κ) coefficient, interpreted according to the criteria established by Landis and Koch [19]. All statistical analyses were performed using SPSS version 26.0 (IBM Corp., Armonk, NY, USA), and a two-sided p-value of < 0.05 was considered statistically significant.

3. RESULTS

3.1. Baseline Patient Characteristics

The final study cohort comprised 63 patients with a mean age of 60.25 ± 10.59 years. The population demonstrated a high prevalence of cardiovascular risk factors: hypertension was present in 58.7% of patients, diabetes mellitus in 38.1%, and a family history of coronary artery disease in 68.3%. Based on the reference

invasive coronary angiography (ICA) results, obstructive coronary artery disease (CAD) was identified in 55.6% of the cohort, reflecting a significant disease burden. Regarding clinical management following ICA, 27 patients (42.9%) underwent percutaneous coronary intervention (PCI), while 8 patients (12.7%) were referred for coronary artery bypass grafting (CABG) (Table 1).

Table 1
Baseline characteristics of the study cohort

Characteristic	Value
Age, years (Mean ± SD)	60.25 ± 10.59
Sex (Male), n (%)	41 (65.1%)
BMI, kg/m□ (Median)	27.12
Hypertension, n (%)	37 (58.7%)
Diabetes Mellitus, n (%)	24 (38.1%)
Family History of CAD, n (%)	43 (68.3%)
No Obstructive CAD (< 50% stenosis), n (%)	28 (44.4%)
Obstructive CAD (≥50% stenosis), n (%)	35 (55.6%)
Medical Therapy, n (%)	28 (44.4%)
PCI, n (%)	27 (42.9%)
CABG, n (%)	8 (12.7%)

3.2. Reliability of Human Expert Consensus

Inter-reader reliability among human observers was initially assessed. For the detection of significant coronary stenosis, inter-observer agreement was moderate across the major epicardial coronary arteries, with Cohen's Kappa (κ) values ranging from 0.477 to 0.519 for the LAD, LCx, and RCA (Table 2). In contrast, agreement for the LMCA was notably poor ($\kappa = 0.096$), reflecting the negligible prevalence of obstructive lesions in this segment and the high sensitivity of this region to image-related artifacts. The assessment's restriction to predefined static frames inherently reduced spatial and contextual information, thereby increasing susceptibility to variability from blooming artifacts and partial volume effects. This limited agreement even among experienced physicians underscores the intrinsic diagnostic challenge of interpreting CCTA solely from static 2D images. Consequently, a third-reader adjudication process was implemented to establish the final expert consensus, which served as the reference standard for the subsequent GP-AI performance analysis.

Table 2
Inter-Observer Agreement (Physician 1 vs. Physician 2)

Vessel	Kappa (p)	ICC (95% CI)	p-value
LMCA	0.096 (0.029)	0.553 (0.354–0.705)	< 0.001
LAD	0.501 (< 0.001)	0.647 (0.475–0.771)	< 0.001
CX	0.519 (< 0.001)	0.622 (0.442–0.754)	< 0.001
RCA	0.477 (< 0.001)	0.790 (0.674–0.868)	< 0.001

3.3. Diagnostic Accuracy for Detection of Significant (≥ 50%) Stenosis

When compared with invasive quantitative coronary angiography (QCA) as the reference standard, the Expert Physician Consensus demonstrated consistently high diagnostic performance across all coronary territories.

Agreement with QCA was robust, with Cohen's Kappa (κ) values ranging from 0.774 to 0.933 (all $p < 0.001$), accompanied by sensitivity and specificity values generally exceeding 90%. These findings indicate that experienced readers maintain high concordance with invasive reference measurements despite the methodological reliance on constrained static image inputs.

In contrast, a marked disparity was observed in the performance of the frontier GP-AI models, which exhibited substantial diagnostic limitations. Notably, none of the GP-AI models achieved statistically significant agreement with QCA in the LAD or LMCA—the most clinically critical coronary segments. For all four models, Kappa values in these territories were negligible and statistically non-significant ($p > 0.05$).

While marginal statistical significance was reached by selected models in the LCx and RCA, overall diagnostic accuracy, including sensitivity and specificity, remained well below the thresholds required for clinical reliability. For instance, in the LAD, Gemini 2.5 achieved a Kappa value of only 0.160 ($p = 0.153$), while ChatGPT-4o demonstrated virtually no agreement ($\kappa = 0.004$, $p = 0.963$). Detailed diagnostic performance metrics for each coronary artery, including positive (PPV) and negative predictive values (NPV), are presented in Table 3.

Table 3
Diagnostic Performance for LMCA-LAD-CX-RCA Stenosis (> 50%) vs. QCA

Evaluator	Kappa (κ)	p-value	Sensitivity	Specificity	PPV	NPV	Accuracy
Gemini 2.5- LMCA	-0.016	0.897	0%	98.4%	0%	98.4%	96.8%
Grok 4- LMCA	0.000	1.000	-	100%	-	100%	100%
Claude 3.5- LMCA	-0.025	0.820	0%	95.1%	0%	98.3%	93.5%
ChatGPT-4o- LMCA	0.000	1.000	-	100%	-	100%	100%
Consensus- LMCA	1.000	< 0.001	100%	100%	100%	100%	100%
Gemini 2.5- LAD	0.160	0.153	78.6%	38.2%	51.2%	68.4%	56.5%
Grok 4- LAD	0.109	0.290	82.1%	29.4%	48.9%	66.7%	53.2%
Claude 3.5- LAD	0.064	0.573	71.4%	35.3%	47.6%	60.0%	51.6%
ChatGPT-4o- LAD	0.004	0.963	85.7%	14.7%	45.3%	55.6%	46.8%
Consensus- LAD	0.774	< 0.001	92.9%	85.3%	83.9%	93.5%	87.1%
Gemini 2.5- CX	0.342	0.007	50.0%	82.9%	58.8%	77.3%	72.1%
Grok 4- CX	0.037	0.771	42.9%	61.0%	36.0%	67.6%	54.8%
Claude 3.5- CX	0.136	0.285	42.9%	70.7%	42.9%	70.7%	61.3%
ChatGPT-4o- CX	0.218	0.060	71.4%	53.7%	44.1%	78.6%	59.7%
Consensus- CX	0.789	< 0.001	90.5%	90.2%	82.6%	94.9%	90.5%
Gemini 2.5- RCA	0.338	0.008	64.0%	70.3%	59.3%	74.3%	67.7%
Grok 4- RCA	0.245	0.042	72.0%	54.1%	51.4%	74.1%	61.3%
Claude 3.5- RCA	0.214	0.090	48.0%	73.0%	54.5%	67.5%	62.9%
ChatGPT-4o- RCA	0.281	0.025	64.0%	64.9%	55.2%	72.7%	64.5%
Consensus- RCA	0.933	< 0.001	96.0%	97.3%	96.0%	97.3%	96.8%

3.4. Continuous Stenosis Assessment and Plaque Characterization

The analysis of continuous stenosis severity revealed a pronounced performance gap between expert readers and the frontier GP-AI models. The Expert Consensus demonstrated high concordance with QCA-derived stenosis degrees across all coronary territories. In contrast, the majority of the GP-AI models failed to show meaningful continuous correlation with invasive reference measurements.

Gemini 2.5 represented a partial exception, achieving a moderate intraclass correlation coefficient (ICC = 0.515, $p < 0.001$) specifically in the right coronary artery (RCA). This finding constituted the sole instance of statistically significant continuous agreement between a GP-AI model and QCA observed in the present study. However, none of the GP-AI models exhibited consistent or clinically acceptable continuous agreement across multiple coronary segments. Detailed ICC results are presented in Table 4.

Regarding plaque characterization, performance across all GP-AI models was uniformly limited. Agreement with the Expert Consensus on plaque morphology (calcified, non-calcified, or mixed) was largely negligible. Only Gemini 2.5 showed slight-to-moderate agreement in the LCx and RCA territories; however, these agreement levels remain insufficient for reliable clinical interpretation. Plaque characterization results are summarized in Table 5.

Table 4
Agreement Between AI Models and QCA-Derived Stenosis Grades

Parameter	Gemini ICC (95% CI)	p	Grok ICC (95% CI)	p	Claude ICC (95% CI)	p	ChatGPT ICC (95% CI)	p	Consensus ICC (95% CI)	p
QCA LMCA grade	0.05 (−0.20 to 0.295)	0.348	0.00 (−0.248 to 0.248)	0.500	0.269 (0.022 to 0.485)	0.017	0.206 (−0.044 to 0.432)	0.053	0.612 (0.43 to 0.747)	< 0.001
QCA LAD grade	0.159 (−0.092 to 0.392)	0.106	0.187 (−0.064 to 0.416)	0.071	0.091 (−0.16 to 0.332)	0.239	0.111 (−0.141 to 0.349)	0.194	0.908 (0.853 to 0.944)	< 0.001
QCA CX grade	0.401 (0.168 to 0.592)	0.001	0.143 (−0.109 to 0.378)	0.131	0.274 (0.027 to 0.488)	0.015	0.141 (−0.111 to 0.375)	0.136	0.847 (0.758 to 0.905)	< 0.001
QCA RCA grade	0.515 (0.306 to 0.677)	< 0.001	0.327 (0.086 to 0.532)	0.004	0.235 (−0.013 to 0.457)	0.032	0.305 (0.062 to 0.514)	0.007	0.942 (0.906 to 0.965)	< 0.001

ICC: Intraclass Correlation Coefficient (95% Confidence Interval)

Table 5
Agreement on Plaque Characterization (Kappa)

Vessel	AI Model	Kappa (κ)	Agreement Level
LMCA	All Models	Non-significant	None
LAD	All Models	Non-significant	None
CX	Gemini 2.5	0.150 ($p = 0.025$)	Slight
RCA	Gemini 2.5	0.203 ($p = 0.009$)	Fair

4. DISCUSSION

This study provides a critical objective clinical evaluation revealing the true diagnostic limitations of general-purpose multimodal artificial intelligence (GP-AI) models in complex cardiovascular imaging tasks. While these models excel in text-based and semantic reasoning, they exhibit significant deficiencies in areas requiring advanced visual-anatomical interpretation, such as coronary CT angiography (CCTA). Although general-purpose artificial intelligence (GP-AI) models are being heavily debated as the next generation of clinical decision support systems [20], our findings emphasize that their visual reasoning capacities are not yet mature enough for complex cardiovascular imaging. Our findings reveal a significant “semantic–visual gap,” defining the fundamental divergence between linguistic intelligence and reliable radiological assessment [21–23].

A critical methodological element in interpreting this study is that both human experts and GP-AI models performed assessments on identical standardized static JPEG images. Neither group was provided with the full volumetric CCTA datasets during the assessment process. Instead, two high-resolution curved MPR frames, intentionally selected to depict the most severe lesions or characteristic anatomy, were used for each major epicardial coronary artery. Thus, the comparison between human and artificial intelligence was performed on a fully symmetrical and methodologically fair basis in terms of data input.

Evaluations were based on two separate reference standards. Plaque morphology was compared with expert physician consensus based on all volumetric CCTA data; the degree of luminal stenosis and the presence of ≥ 50% obstructive disease were analyzed based on quantitative coronary angiography (QCA) results. This dual-reference approach provides a sound validation basis for the biological and technical nature of plaque characterization and hemodynamically significant stenosis detection [24–26].

The significant performance difference between GP-AI models and human experts is largely related to the diagnostic challenges inherent in the static two-dimensional imaging approach. The moderate-to-low inter-observer agreement observed among human experts—specifically the markedly low agreement in the LMCA ($\kappa = 0.096$) highlights the intrinsic difficulty of this task. This low concordance in the LMCA is likely due to its short anatomical length, the relatively lower prevalence of isolated obstructive disease in this segment, and its high susceptibility to artifacts in static frames. These factors create a 'diagnostic blind spot' when volumetric scrolling is unavailable, explicitly demonstrating that static-image CCTA assessment is inherently constrained, even for highly trained specialists.

Consequently, it is entirely unsurprising that GP-AI models, which lack both contextual clinical knowledge and full three-dimensional spatial reasoning, failed significantly in the LMCA and LAD. These 'small-target' segments require advanced anatomical contextual inference that current multimodal architectures lack. The models' inability to achieve statistical significance in these prognostically critical territories—often misinterpreting vascular overlaps, blooming artifacts, or calcium shadows as stenosis—represents a major barrier to their clinical reliability and increases the risk of severe misinterpretations [27–30].

Within this general picture of failure, the limited performance improvement observed in the right coronary artery (RCA) constitutes a notable exception. The more isolated anatomical course and relatively lower artifact load of the RCA may have allowed some models, particularly Gemini 2.5, to partially capture the change in stenosis severity on a continuous scale. However, the discrepancy between continuous correlation and binary classification success reveals that the models struggle to define clinically meaningful decision limits and exhibit a tendency toward systematic overestimation.

Finally, while the GP-AI models used (GPT-4o, Gemini 2.5, Claude 3.5 Sonnet, and Grok 4) were the most current versions as of February 2026, the rapid evolution of AI technologies means that more advanced versions are quickly introduced. However, since this study aimed to reveal the fundamental competency

limitations of general-purpose architectures in tasks requiring expert-level radiological evaluation, the findings retain their conceptual and methodological validity.

Limitations

Our study design carries specific inherent limitations that must be explicitly acknowledged. First, the unblinded selection of key frames by the preparation team introduces a potential selection bias, as it intentionally guides the presented images toward the most severe lesion. While this does not reflect real-world volumetric CCTA reading where the clinician must actively search for pathology, it was a necessary methodological compromise to establish a standardized ground truth for evaluating pure visual diagnostic capacity. Second, because this study evaluates decision-making within a constrained framework utilizing only a few standardized static images per vessel rather than actual clinical volumetric interpretation, extrapolation of these findings to real-world, unconstrained clinical practice should be interpreted with caution.

5. CONCLUSION

This study demonstrates that consensus interpretation generated by experienced physicians exhibits strong agreement with invasive quantitative coronary angiography (QCA) when using a ≥ %50 stenosis threshold and remains the most reliable reference approach in coronary CT angiography (CCTA) assessment.

In contrast, current general-purpose multimodal artificial intelligence (GP-AI) models have shown significant limitations in terms of visual diagnostic performance. Despite their known competencies in text-based medical tasks, these systems have failed to achieve a consistent and clinically acceptable level of accuracy when applied to static two-dimensional CCTA images, particularly in anatomically and prognostically critical coronary segments. These findings indicate that direct transfer of general visual reasoning capabilities to highly specialized cardiovascular imaging tasks is not feasible under current conditions.

Therefore, at the current stage, it is not medically appropriate for GP-AI systems to replace physicians in interpreting coronary CT angiography results or for patients to directly consult these systems for diagnostic decisions. CCTA assessment currently requires the clinical experience, anatomical knowledge, and contextual interpretation skills of expert cardiologists and radiologists.

The limited performance signal observed with the Gemini 2.5 model in the right coronary artery (RCA) suggests that clinically relevant feature recognition capacity may be partially present in some pioneering architectures. However, this finding is insufficient to support the independent clinical use of GP-AI models, suggesting that the identified shortcomings are related to a lack of domain adaptation and input representation.

In conclusion, current GP-AI models are not suitable for the independent clinical interpretation of coronary CT angiography. Importantly, this limitation applies specifically to broad, non-domain-specific architectures and does not detract from the proven clinical value of dedicated, task-specific cardiovascular AI platforms. Their potential as supportive tools in the future can only be achieved through the development of architectures incorporating domain-specific targeted training, radiological physics, and three-dimensional spatial relationships, and their use under expert physician supervision.

Declarations

Ethics approval and consent to participate: This study was approved by the Baskent University Institutional Review Board (Project no: KA25/407). All procedures were performed in accordance with the principles of the Declaration of Helsinki. Given the retrospective nature of the study, the requirement for written informed consent was waived by the Institutional Review Board.

Consent for publication: Not applicable.

Availability of data and materials: The datasets generated and/or analysed during the current study are not publicly available due to patient privacy and institutional regulations but are available from the corresponding author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: This study was supported by the Baskent University Research Fund (Project no: KA25/407).

Authors' contributions: S.O. conceptualized and designed the study. S.O., İ.G.B., and G.E. collected the clinical data and performed the expert image evaluations. Z.G.B. and S.O. conducted the artificial intelligence model assessments and statistical analysis. S.O. wrote the main manuscript text. M.Y. supervised the research project and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements: Not applicable.

References

1. Vrints CJM, Senior R, Crea F, et al. 2024 ESC Guidelines for the management of chronic coronary syndromes. Eur Heart J. 2024;45(36):3415-3537.
2. Kelion AD, Nicol ED. The rationale for the primacy of coronary CT angiography in the National Institute for Health and Care Excellence (NICE) guideline (CG95) for the investigation of chest pain of recent onset. J Cardiovasc Comput Tomogr. 2018;12:516–22.
3. Abbara S, Blanke P, Maroules CD, et al. SCCT guidelines for the performance and acquisition of coronary computed tomographic angiography: A report of the society of Cardiovascular Computed Tomography Guidelines Committee. J Cardiovasc Comput Tomogr. 2016;10(6):435-449.
4. Conte E, Sala E. AI-assisted CCTA: supporting diagnosis across the CAD spectrum. Int J Cardiovasc Imaging. 2025;41:825-826.
5. D'Costa Z, Karlsberg RP, Cho GW. Artificial-intelligence-assisted CCTA quantifies sex differences in coronary atherosclerotic burden at low atheroma volumes. IJC Heart & Vasculature. 2025;60:101758.
6. Liao J, Huang L, Qu M, Chen B, Wang G. Artificial Intelligence in Coronary CT Angiography: Current Status and Future Prospects. Front Cardiovasc Med. 2022;9:896366.
7. Bozyel S, Şimşek E, Koçyiğit Burunkaya D, et al. Artificial intelligence-based clinical decision support systems in cardiovascular diseases. Anatol J Cardiol. 2024;28(2):74-86.
8. van Assen M, De Cecco CN, et al. Inter-observer variability of coronary artery calcium scoring and CCTA interpretation. J Cardiovasc Comput Tomogr. 2019;13(4):228–233.
9. Budoff MJ, et al. Interobserver variability among expert readers quantifying plaque volume on coronary CT angiography. J Cardiovasc Comput Tomogr. 2022;16(6):501–507.
10. Interobserver variability of coronary stenosis characterization and its relation to plaque composition. JACC Cardiovasc Imaging. 2024.
11. Chen M, Wang X, Hao G, et al. Diagnostic performance of deep learning-based vascular extraction and stenosis detection technique for coronary artery disease. Br J Radiol. 2020;93:20191028.
12. CathAI: fully automated coronary angiography interpretation and stenosis estimation. NPJ Digit Med. 2023;6(1):142.
13. Sarangi PK, Datta S, Panda BB, et al. Evaluating ChatGPT-4's Performance in Identifying Radiological Anatomy in FRCR Part 1 Examination Questions. Indian J Radiol Imaging. 2025;35:287-294.
14. Gillette J, Lu M, Heston TF. Large Language Models Perform at Chance Level in the Diagnosis of Pediatric Pneumonia Using Chest Radiographs. Cureus. 2025;17(9):e92596.

15. Gemini Team, Google. Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805. 2023.
16. Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical Report. 2024.
17. OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. 2023.
18. Reiber JH, Serruys PW, Kooijman CJ, et al. Assessment of short-, medium-, and long-term variations in arterial dimensions from computer-assisted quantitation of coronary cineangiograms. Circulation. 1985;71(2):280-288.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174.
20. Güneş YC, Cesur T. Large Language Models: Could They Be the Next Generation of Clinical Decision Support Systems in Cardiovascular Diseases? Anatol J Cardiol. 2024;28(7):371-372.
21. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
22. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–98.
23. Large Language Models for disease diagnosis: a scoping review. NPJ Digit Med. 2025;8:11.
24. Hulten E, Villines TC, Cheezum MK, et al. The role of coronary CT angiography in the diagnosis and management of coronary artery disease. J Nucl Cardiol. 2017;24(5):1609–1624.
25. Knuuti J, Wijns W, Saraste A, et al. 2019 ESC Guidelines for the diagnosis and management of chronic coronary syndromes. Eur Heart J. 2020;41(3):407–477.
26. Clinical expert consensus document on quantitative coronary angiography. Cardiovasc Interv Ther. 2020;35(2):105–116.
27. Park SH. Artificial intelligence in radiology: practical issues and challenges. Radiology. 2018;287(3):749–772.
28. Large Language Models in Medical Image Analysis: A Systematic Review. Bioengineering (Basel). 2025;12(8):818.
29. Multimodal Large Language Models in Medical Imaging. Korean J Radiol. 2025;26(9):843–853.
30. Best Practices for the Safe Use of Large Language Models and Generative AI in Radiology. Radiology. 2025;312(3):e241516.

Figures

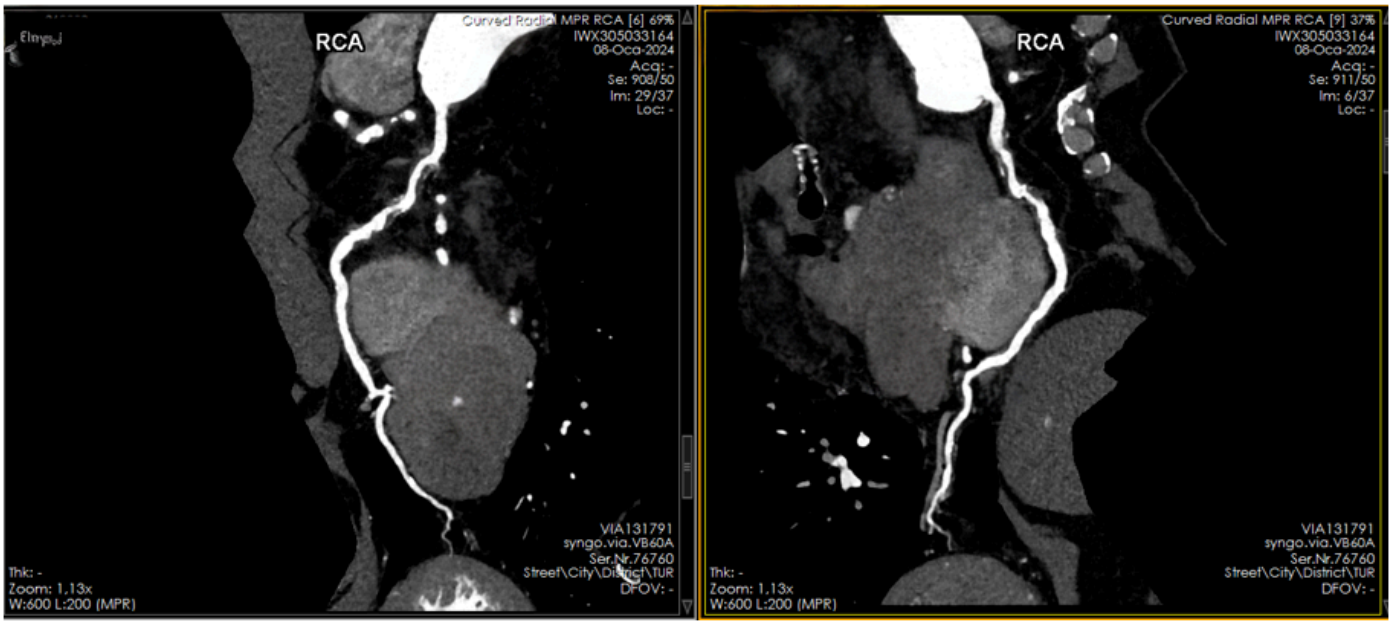


Figure 1

Study Methodology for Artificial Intelligence Assessment

An example of a Curved Multi-Planar Reformat (Curved MPR) image of the Right Coronary Artery (RCA) from two standardized sequences. This technique straightens the tortuous vessel along the centerline to clearly visualize the vessel trajectory and plaque burden and facilitate lumen assessment.

JSON

```
{
  "Task": "Evaluate Coronary CT Angiography for Significant Stenosis (>50%)",
  "Input_Data": ["Axial MPR", "Curved MPR", "3D Volume Rendering"],
  "Response_Format": {
    "Segment": "[e.g., LAD Proximal]",
    "Stenosis_Present": "Boolean (true/false)",
    "Stenosis_Percentage": "[0%, 1-24%, 25-49%, 50-69%, 70-99%, 100%]",
    "Plaque_Type": "[Calcified, Non-calcified, Mixed, None]",
    "Confidence": "[High, Medium, Low]"
  }
}
```

Figure 2

Standard JSON Prompt Template Used to Query Artificial Intelligence Models.

This structured command set was sent to all artificial intelligence models evaluated in the study. This template enabled the models to analyze visual data and report the presence of significant stenosis (≥50%), estimated stenosis percentage, plaque characterization, and confidence score in a standardized format.

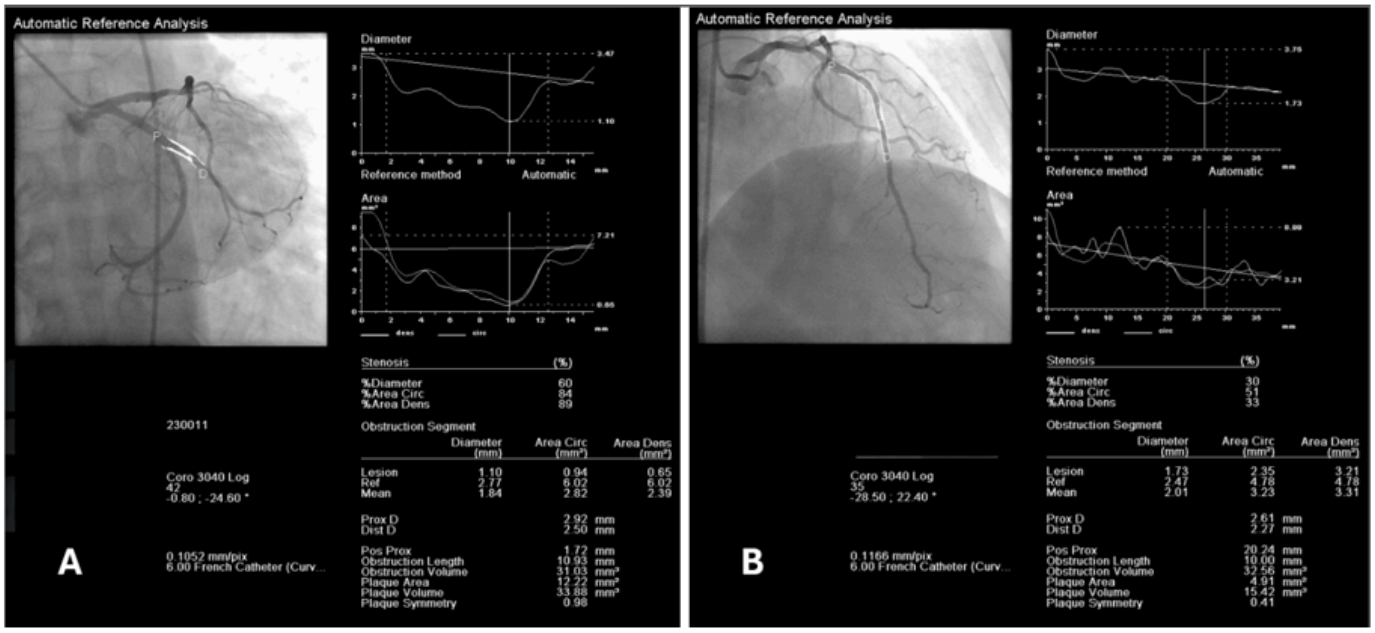


Figure 3

Quantitative Coronary Angiography (QCA) Analysis (Gold Standard). Examples of analyses performed with syngo QCA software (Siemens Healthineers), used as the reference method in the study. (A) Example of significant stenosis in the Left Anterior Descending Artery (LAD). The 60% diameter stenosis measured in the analysis meets the study's significant stenosis threshold (≥50%). (B) Example of non-obstructive disease in the Right Coronary Artery (RCA). The 30% diameter stenosis measured in the analysis was not considered significant.