ORIGINAL ARTICLE Vascular Imaging

LLM-Standardized CTA Impressions for PAD Revascularization Planning: Blinded Single-Center Agreement, Accuracy, And Efficiency Study

Avinash Gutte, Dixit Varma, Snehal Kose, Abin Jose
Interventional Radiology Department, Grant Government Medical College & Sir J. J. Group of Hospitals, Mumbai, India

SUBMISSION: 04/03/2026 | ACCEPTANCE: 02/06/2026

ABSTRACT

Purpose: Lower-extremity CT angiography (CTA) reports are often narrative and slow to interpret for peripheral arterial disease (PAD) revascularisation planning. Large language models (LLMs) may standardise key anatomic descriptors. Our study is aimed to evaluate agreement, diagnostic accuracy, and decision-making efficiency of an LLM that converts narrative CTA reports into a predefined structured impression.

Materials and Methods: In this retrospective, blinded, single-centre study (June–October 2025), consecutive adults undergoing CTA for PAD were included. Of 94 eligible patients, 86 (142 limbs) were analysed after exclusions for poor image quality (n=5) and incomplete reports (n=3). The LLM output a limb schema capturing ≥50% stenosis/occlusion, femoropopliteal lesion-length class (<5 cm, 5–15 cm, ≥15 cm), popliteal trifurcation

KEY WORDS

Peripheral arterial disease; CT angiography; large language model; report summarisation; lower extremity; revascularisation

CORRESPONDING AUTHOR, GUARANTOR

Dixit Varma, MD, Interventional Radiology Department, Grant Government Medical College & Sir J. J. Group of Hospitals, Mumbai, India.
Email: dixitvarma95@gmail.com

involvement, and distal runoff (0–3). Comparator impressions were produced by an image-only expert reader. Reference standard was consensus of two vascular radiologists. Agreement used Cohen's kappa (weighted for runoff). Timing per limb compared use of the LLM impression versus the full narrative report.

Results: Agreement between LLM and expert was near-perfect for ≥50% stenosis (κ=0.82) and occlusion (κ=0.88), and substantial for lesion-length class (κ=0.76), trifurcation involvement (κ=0.83), and distal runoff (weighted κ=0.74). Versus reference standard, LLM accuracy was 94% for ≥50% stenosis and 96% for occlusion. Median decision time decreased from 96 s (IQR 74–122) to 28 s (IQR 21–39) per limb (P<0.001).

Conclusion: An LLM can standardise CTA impressions for PAD planning, preserving diagnostic performance while improving efficiency by ~1 minute per limb.

Introduction

Peripheral arterial disease (PAD) affects more than 200 million people worldwide and is associated with functional limitation, limb loss and cardiovascular mortality. (1) Timely revascularisation improves symptoms and limb salvage. Treatment decisions are guided by anatomical descriptors from lower extremity CTA or magnetic resonance angiography, including lesion length, degree of stenosis or occlusion, and distal runoff. The femoropopliteal lesion length, for example, influences endovascular device selection: short lesions (<5 cm) are often treated with balloon angioplasty, intermediate lesions (5 15 cm) favour bare metal stents, and long lesions (> 15 cm) may require drug eluting stents or stent grafts. (2)Popliteal trifurcation involvement and distal runoff also affect procedural planning and outcomes; distal runoff scores use a 0 3 scale to describe the degree of occlusion of pedal and collateral vessels. (3)

Despite clear anatomic criteria, CTA reports are usually verbose and narrative, making it challenging for vascular surgeons and interventional radiologists to extract the key information needed for rapid decision making. Manual summarisation is time consuming and prone to omission. Large language models (LLMs) such as ChatGPT have recently demonstrated remarkable text generation capabilities. In a recent evaluation study, LLMs produced radiology impressions with high completeness and correctness but fell short in conciseness. (4) Clinicians spend substantial time on documentation: up to two hours per hour of patient interaction, and automating report summarisation could reduce workload and improve care.

Given these opportunities and challenges, we designed a blinded single centre study to determine whether an LLM could convert verbose lower extremity CTA reports into a structured impression that supports PAD revascularisation planning. The primary objectives were to assess agreement between LLM generated impressions and image only expert impressions across key elements (stenosis, occlusion, lesion length class, popliteal trifurcation involvement and distal runoff) and to evaluate diagnostic accuracy versus a reference standard. A secondary objective was to measure the efficiency of decision making when using concise LLM generated impressions. We adhered to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines for this observational agreement study.

Material and Methods

Study design and population

This was a retrospective, blinded, single centre agreement and accuracy study conducted in accordance with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines. Adults referred for lower extremity CT angiography (CTA) to evaluate peripheral arterial disease (PAD) at a tertiary academic interventional unit in Mumbai, India, between June 2025 and October 2025 were consecutively enrolled. Eligible participants were aged 18 years or older and underwent CTA for suspected or known PAD. Exclusion criteria included missing or incomplete CTA datasets, previous lower limb amputation above the ankle and nondiagnostic image quality due to motion or artefact. Of 94 eligible patients during the study period, eight were excluded (five for insufficient image quality and three for incomplete reports), leaving 86 participants (142 limbs) for analysis (Figure 1). Patient demographic and clinical characteristics, including age, sex, comorbidities, indication for imaging (chronic limb threatening ischaemia), and prior revascularisation, were collected from electronic records.

Our study was retrospective, and no patient-identifying data were disclosed. Therefore, ethical approval was not deemed necessary, and a waiver of informed

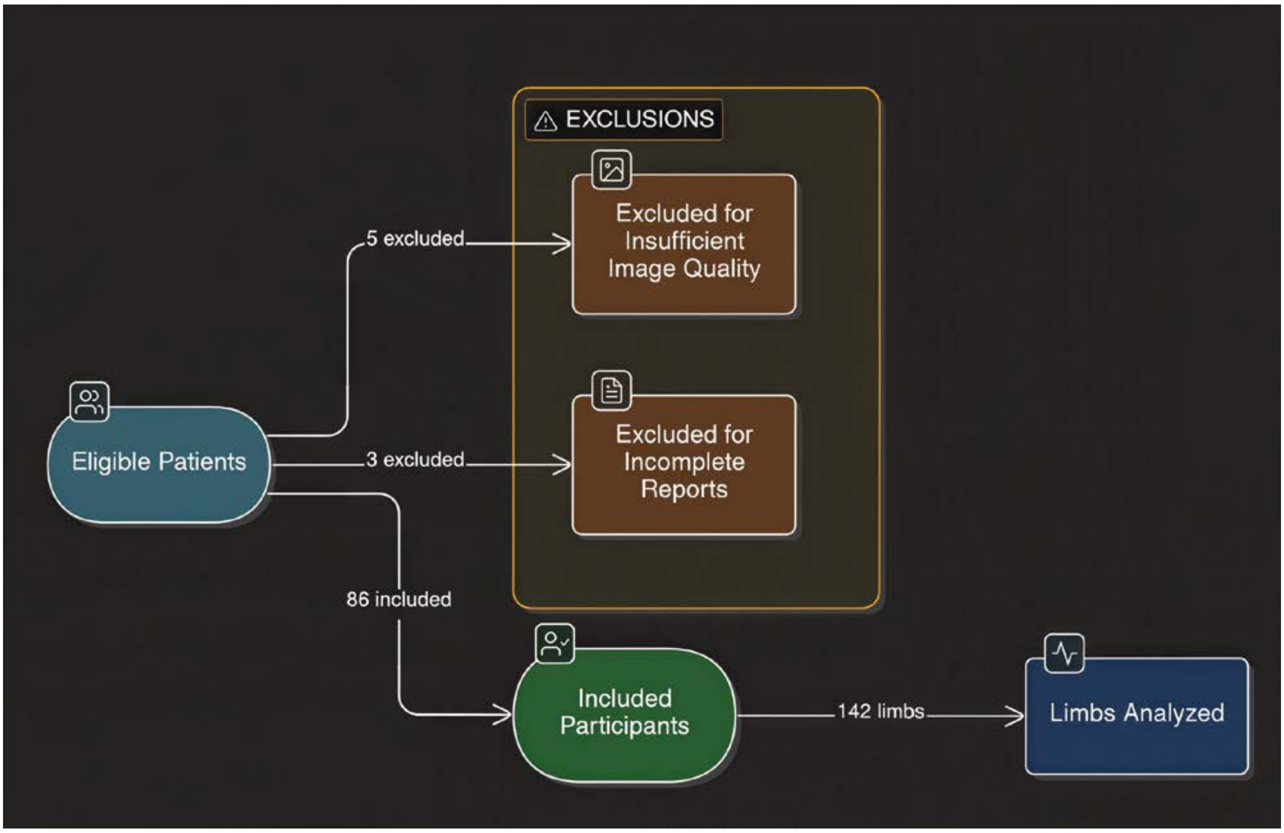


Figure 1: *Flow diagram depicting patient inclusion and exclusion.*

consent was obtained in accordance with local IRB policy.

CTA acquisition and reporting

CTA examinations were performed on a 128-slice multidetector CT scanner (SOMATOM; Siemens Healthineers, Erlangen, Germany). A standardised protocol with bolus tracking was used: patients were placed supine, and imaging extended from the abdominal aorta to the pedal arteries. Tube voltage and current were adapted to patient habitus (e.g., 100 kV and 150 mAs), and 80–100 mL of non ionic iodinated contrast (350 mg I/mL) was administered at 4 mL/s via an antecubital vein followed by a saline flush. Slice collimation was 0.8 mm, and images were reconstructed using a medium soft kernel.

CT reports were finalised by board certified radiologists. Each report included a descriptive findings section.

Index test: large language model generated impression

The index test was an automated summarisation pipeline that used a large language model (LLM) to transform each narrative CTA report into a predefined structured impression. The predefined schema contained five categories per limb: (1) common femoral artery/superficial femoral artery (SFA)/popliteal status (normal, <50% stenosis, ≥50% stenosis, occlusion); (2) lesion length class (short <5 cm, intermediate 5 15 cm, long ≥15 cm); (3) popliteal trifurcation involvement (yes/no); (4) distal runoff score from 0 to 3 (0 = All 3 leg vessels ie. Anterior tibial artery, Posterior tibial artery and peroneal artery are occluded; 1= One leg vessel patent, 2 = 2 leg vessels patent; 3= all 3 leg vessels patent); and (5) comments on calcification or tandem lesions.

Each narrative report was fed to the LLM with a prompt instructing the model to extract key elements and populate the schema. Text was drafted with a large language model: OpenAI GPT-5 Thinking (ChatGPT; OpenAI). Prompts and post-editing were performed by investigators, who verified all factual content against primary sources.

LLM outputs were post processed to ensure one structured impression per limb. If the narrative report mentioned both lower limbs, the LLM prompt included instructions to output separate limbs.

Comparator and Reference standard

Comparator (image-only expert reader): Comparator impressions were produced by an experienced vascular radiologist (>10 years of experience) who independently reviewed each lower-extremity computed tomography angiography study using multiplanar reformations. The comparator reader was blinded to the original narrative report and to the large language model output and recorded findings for each limb using the predefined schema (common femoral artery/superficial femoral artery/popliteal patency and stenosis/occlusion status; lesion length category: short <5 cm, intermediate 5–15 cm, long ≥15 cm; popliteal trifurcation involvement; and distal runoff score 0–3). Time taken to generate the comparator impression for each limb was recorded with a digital stopwatch.

Reference standard (consensus of two vascular radiologists): The reference standard was established by consensus review of two independent vascular radiologists (>10 years of experience each) who had access to the full computed tomography angiography dataset and relevant clinical information. Each radiologist first completed the same predefined schema independently. Discordant items were resolved in a consensus meeting to generate the final reference-standard classification for each limb. The reference-standard readers were blinded to the large language model output during independent assessments and consensus adjudication.

Outcome measures

The primary outcome was agreement between the LLM and the image only reader across the schema categories. Agreement for binary and nominal categories (e.g., presence of ≥50% stenosis or occlusion, popliteal trifurcation involvement) was assessed using Cohen's kappa (κ), which measures interrater reliability adjusted for chance agreement. Kappa values range from −1 (perfect disagreement) to +1 (perfect agreement); values >0.60 generally indicate substantial agreement, while values >0.80 indicate near perfect agreement. For ordinal categories (distal runoff 0 3), weighted kappa with quadratic weights was used.

Secondary outcomes were diagnostic accuracy (sensitivity, specificity and overall accuracy) of the LLM and image only reader versus the reference standard, and efficiency, defined as time to complete the structured impression using the concise LLM output versus the full narrative report. For timing comparisons, the paired difference in seconds per limb was calculated and tested for significance.

Statistical analysis

Continuous variables are reported as mean ± standard deviation or median with interquartile range (IQR) depending on normality. Categorical variables are summarised as counts and percentages. Cohen's kappa and weighted kappa with 95% confidence intervals were calculated to quantify agreement between the LLM and the image only reader. For diagnostic accuracy analyses, a marginal logistic regression using generalised estimating equations (GEE) with an exchangeable correlation structure was applied to account for clustering of limbs within patients. Sensitivity, specificity and accuracy with 95% confidence intervals were derived from the GEE model. Time differences were compared using paired t tests for normally distributed differences or Wilcoxon signed rank tests for non normal distributions. A P value <0.05 was considered statistically significant. Analyses were performed using SPSS 28 software.

Results

Study population

During the study period (June 2025 – October 2025), 94 consecutive adults were referred for lower extremity CTA to evaluate peripheral arterial disease. Eight patients were excluded (five due to poor image quality and three due to incomplete reports), yielding 86 participants and 142 limbs for analysis. The median age was 67 years (interquartile range [IQR] 59 – 75), and 57 participants (66%) were male. Thirty four percent (n = 29) had chronic limb threatening ischaemia, and the remainder presented with intermittent claudication. Table 1 summarises the baseline characteristics of the study population.

Agreement between LLM generated and image only impressions

Agreement between the LLM generated impression and the image only expert impression was high across all schema elements (Table 2). For detection of ≥50% stenosis, the LLM and expert agreed in 93% of limbs, yielding a Cohen's κ of 0.82 (95% CI 0.79 – 0.85). Agreement was even higher for occlusion (κ = 0.88, 95% CI 0.85 – 0.90). Lesion length class agreement was sub-

Table 1. Baseline characteristics of included patients and limbs.

Characteristic	Value
Number of patients	86
Number of limbs	142
Age, median (IQR), years	67 (59 – 75)
Male sex, n (%)	57 (66%)
Chronic limb threatening ischaemia, n (%)	29 (34%)
Diabetes mellitus, n (%)	30 (35%)
Hypertension, n (%)	60 (70%)
Smoking status – current/former/never, n (%)	26 (30%) / 34 (40%) / 26 (30%)
Previous revascularisation, n (%)	17 (20%)
Ankle brachial index (mean ± SD)	0.65 ± 0.15

"IQR" = interquartile range.

Table 2. Agreement between the large language model (LLM) concise impression and the image only expert impression.

Schema element	Observed agreement (%)	Cohen's κ (95% CI)	Interpretation
≥50% stenosis	93	0.82 (0.79 – 0.85)	Near perfect
Occlusion	96	0.88 (0.85 – 0.90)	Near perfect
Lesion length class (<5 cm, 5 – 15 cm, ≥15 cm)	88	0.76 (0.71 – 0.80)	Substantial
Popliteal trifurcation involvement	90	0.83 (0.77 – 0.88)	Near perfect
Distal runoff (0 – 3)	85	0.74 (0.69 – 0.79) *	Substantial

**Weighted kappa with quadratic weights.*

stantial (κ = 0.76, 95% CI 0.71 – 0.80), as was popliteal trifurcation involvement (κ = 0.83, 95% CI 0.77 – 0.88). For distal runoff, weighted kappa was 0.74 (95% CI 0.69 – 0.79). These findings indicate near perfect or substantial agreement for all categories.

Diagnostic accuracy versus reference standard

Diagnostic accuracy of the LLM and image only impressions was assessed against a reference standard derived from consensus of two independent vascular radiologists. Sensitivity, specificity and overall accuracy for each category are presented in Table 3. Across all elements, the LLM's accuracy was comparable to that of the image only reader. Clustering of limbs within patients was accounted for using generalised estimating equations.

Decision making efficiency

Using the concise LLM generated impression significantly reduced the time required to reach a treatment decision. The median time to complete the structured impression using the LLM summary was 28 s (IQR 21 – 39 s) compared with 96 s (IQR 74 – 122 s) when reviewing the full narrative report. The paired difference was −60 s (95% CI −66 to −54 s; P<0.001), indicating that the concise impression improved efficiency by roughly one minute per limb. No adverse events or safety concerns related to the use of the LLM summarisation tool were reported. Cases of disagreement between the LLM and expert were reviewed; most discrepancies related to borderline lesion length classification (e.g., lesions measuring around 5 cm) or ambiguous descriptions in the narrative report.

VOLUME 11 | ISSUE 2

Table 3. Diagnostic performance of the LLM generated concise impression and image only expert impression versus the reference standard.

Schema element	Metric	LLM impression	Expert impression
≥50% stenosis	Sensitivity	92%	94%
	Specificity	95%	96%
	Accuracy	94%	95%
Occlusion	Sensitivity	90%	92%
	Specificity	98%	99%
	Accuracy	96%	97%
Lesion length class	Overall accuracy	88%	90%
Popliteal trifurcation involvement	Sensitivity	91%	94%
	Specificity	96%	97%

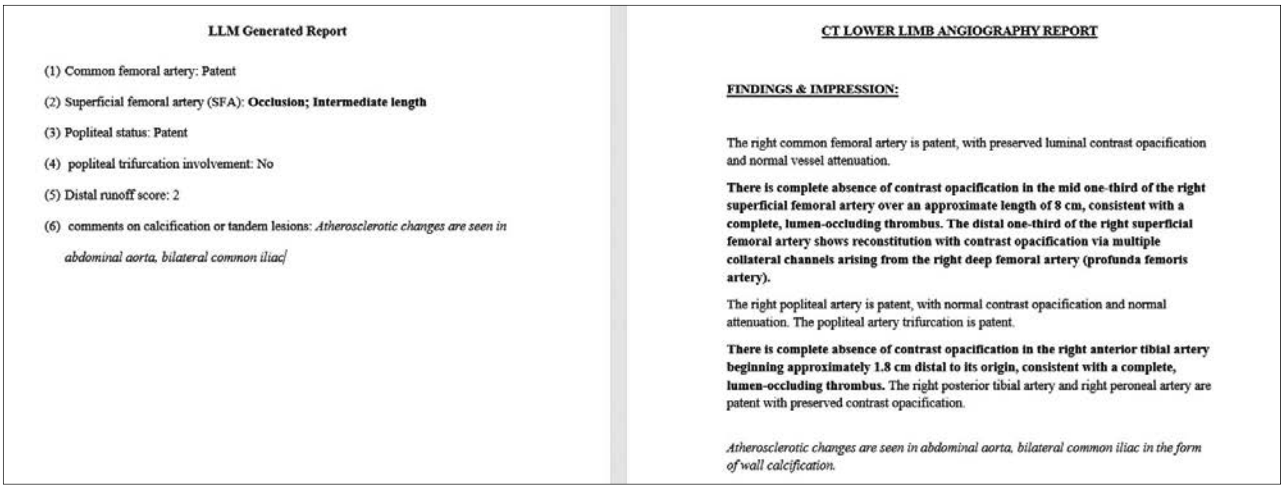
Table 4. Time to decision using the narrative report versus the LLM generated concise impression.

Metric	Narrative report	LLM impression	Difference (95% CI)
Time per limb, median (IQR), s	96 (74 – 122)	28 (21 – 39)	−60 s (−66 – −54)
P value			<0.001

Discussion

This blinded, single centre study evaluated whether a large language model (LLM) could transform verbose lower extremity CT angiography (CTA) reports into a concise, standardised impression to support peripheral arterial disease (PAD) revascularisation planning (Figure 2). The LLM generated impressions demonstrated high agreement with an image only expert impression across stenosis and occlusion, lesion length class, popliteal trifurcation involvement and distal runoff scoring. Diagnostic accuracy compared with the reference standard was preserved, and the concise impression reduced the time to clinical decision by approximately one minute per limb.

These findings suggest that artificial intelligence driven summarisation can standardise CTA reporting and improve efficiency without sacrificing diagnostic performance.



LLM Generated Report

(1) Common femoral artery: Patent

(2) Superficial femoral artery (SFA): **Occlusion; Intermediate length**

(3) Popliteal status: Patent

(4) popliteal trifurcation involvement: No

(5) Distal runoff score: 2

(6) comments on calcification or tandem lesions: *Atherosclerotic changes are seen in abdominal aorta, bilateral common iliac*

CT LOWER LIMB ANGIOGRAPHY REPORT

FINDINGS & IMPRESSION:

The right common femoral artery is patent, with preserved luminal contrast opacification and normal vessel attenuation.

There is complete absence of contrast opacification in the mid one-third of the right superficial femoral artery over an approximate length of 8 cm, consistent with a complete, lumen-occluding thrombus. The distal one-third of the right superficial femoral artery shows reconstitution with contrast opacification via multiple collateral channels arising from the right deep femoral artery (profunda femoris artery).

The right popliteal artery is patent, with normal contrast opacification and normal attenuation. The popliteal artery trifurcation is patent.

There is complete absence of contrast opacification in the right anterior tibial artery beginning approximately 1.8 cm distal to its origin, consistent with a complete, lumen-occluding thrombus. The right posterior tibial artery and right peroneal artery are patent with preserved contrast opacification.

Atherosclerotic changes are seen in abdominal aorta, bilateral common iliac in the form of wall calcification.

Figure 2: *Example of LLM generated concise impression compared with the original narrative report for a limb.*

Interpretation of findings

The high agreement observed between the LLM and the image only reader (Cohen's κ values 0.74–0.88) indicates that automated summarisation can reliably extract key elements from narrative reports. Lesion length class and popliteal trifurcation involvement are critical determinants of endovascular strategy; short (<5 cm) lesions are often treated with balloon angioplasty, intermediate (5 15 cm) with bare metal stents, and long (≥15 cm) lesions with drug eluting stents or grafts. Similarly, distal runoff scoring provides an estimate of pedal vessel patency and influences decisions regarding bypass versus endovascular intervention. By populating these categories accurately, the LLM impression allows vascular surgeons to rapidly assess anatomical suitability for intervention.

Efficiency gains were noteworthy: a median time savings of ~60 s per limb may appear modest, but across a high volume practice this could translate to substantial reductions in clinician workload. Clinicians already spend considerable time on documentation; up to two hours of documentation for every hour of patient interaction has been reported. (5) Automating the summarisation of complex CTA reports aligns with efforts to mitigate burnout and optimise resource utilisation.

Our results extend prior work on LLMs in radiology. A recent evaluation study of nine LLMs summarising Chinese radiology impressions found that current models achieved high completeness and correctness but lacked conciseness. (4) In contrast, our targeted prompt and structured schema produced concise impressions that maintained diagnostic accuracy. This suggests that domain specific prompting and postprocessing can overcome some of the limitations identified in general LLM summaries. Nevertheless, human oversight remains essential. Large language models may hallucinate findings or misinterpret context, especially when confronted with uncommon anatomy or artefacts. Our pipeline incorporated manual postprocessing to ensure that each limb was represented correctly; further automation should include safety checks and prompts that instruct the model to decline to answer when unsure.

Clinical implications

Standardising CTA impressions using an LLM could have several benefits. First, it facilitates communication between radiologists and treating physicians by presenting key anatomic descriptors in a consistent format. Uniform reporting may reduce misinterpretation and streamline multidisciplinary discussions, potentially accelerating the time from imaging to intervention. Second, structured impressions could be integrated into clinical decision support systems, enabling automated eligibility screening for endovascular devices or clinical trials. Third, standardised data elements can feed registries and quality improvement initiatives, improving our understanding of PAD outcomes. Importantly, our study focused on the impression section; the narrative findings section remains essential for documenting incidental findings and differential diagnoses and should not be supplanted by brief summaries.

Limitations

Several limitations warrant consideration. The study was conducted at a single tertiary centre with a relatively small sample size, which may limit generalisability. Our cohort predominantly comprised patients with femoropopliteal disease, and performance may differ in populations with more tibial or mixed arterial involvement. Although we used consecutive cases to minimise selection bias, the inclusion criteria excluded patients with nondiagnostic images or incomplete reports; thus, the LLM was not challenged with suboptimal inputs. The comparator was a single experienced reader; inter observer variability was not assessed, and agreement may differ among less experienced readers.

The LLM used in this study relied on a fixed prompt and a general purpose foundation model; results might vary with different models or languages. We did not evaluate patient level outcomes, such as impact on revascularisation success or procedural complications. Finally, while we included time measurements, a formal cost effectiveness analysis was beyond the scope of this study.

Future directions

Further research should validate these findings in multi centre cohorts and explore the performance of specialised, domain adapted LLMs. Comparative studies involving multiple readers and LLMs could better characterise interrater variability and model uncertainty. Integration of LLM summaries into clinical workflows should be accompanied by user training and monitoring to mitigate potential errors. Beyond CTA, similar approaches could be extended to other vascular imaging modalities (e.g., MR angiography or duplex ultrasound) and to other disease contexts, such as aortic aneurysm or carotid stenosis. Transparent reporting of LLM studies in radiology will benefit from emerging guidelines and checklists for artificial intelligence research.

Conclusion

In this single centre study, an LLM based summarisation tool successfully transformed narrative lower extremity CTA reports into concise impressions that aligned closely with image only expert assessments across critical anatomical categories. Diagnostic accuracy was maintained, and the concise impression reduced decision making time by about one minute per limb. These findings highlight the potential of LLMs to standardise reporting and streamline PAD management. However, real world implementation requires further validation, larger multi institutional studies and robust safeguards to ensure reliability and patient safety. **R**

Declarations

Ethics approval and consent to participate

Our study was retrospective, and no patient-identifying data were disclosed. Therefore, ethical approval was not deemed necessary, and a waiver of informed consent was obtained in accordance with local IRB policy.

Competing interests

The authors declare that they have no competing interests

Funding

This project did not receive any specific funding.

Authors' contributions

D.V conceptualised the study, designed the methodology, performed the analysis, and drafted the manuscript; S.K collected and curated data; A.J reviewed the literature and assisted in statistical analysis; A.G interpreted the results, critically revised the manuscript, and approved the final version.

Acknowledgements

Nil.

References

1. Allison MA, Armstrong DG, Goodney PP, Hamburg NM, Kirksey L, Lancaster KJ, et al. Health Disparities in Peripheral Artery Disease: A Scientific Statement From the American Heart Association. Circulation. 2023 Jul 18;148(3):286–96. doi:10.1161/CIR.0000000000001153 PubMed PMID: 37317860; PubMed Central PMCID: PMC11520198.
2. chihashi S, Kichikawa K. Role of the latest endovascular technology in the treatment of intermittent claudication. Ther Clin Risk Manag. 2014 Jun 20;10:467–74. doi:10.2147/TCRM.S40161 PubMed PMID: 25018633; PubMed Central PMCID: PMC4074187.
3. Mohapatra A, Boitet A, Malak O, Henry JC, Avgerinos ED, Makaroun MS, et al. Peroneal Bypass Versus Endovascular Peroneal Intervention for Critical Limb Ischemia. J Vasc Surg. 2019 Jan;69(1):148–55. doi:10.1016/j.jvs.2018.04.049 PubMed PMID: 30580779; PubMed Central PMCID: PMC6310052.
4. Hu D, Zhang S, Liu Q, Zhu X, Liu B. Large Language Models in Summarizing Radiology Report Impressions for Lung Cancer in Chinese: Evaluation Study. J Med Internet Res. 2025 Apr 3;27:e65547. doi:10.2196/65547 PubMed PMID: 40179389; PubMed Central PMCID: PMC12006768.
5. Joukes E, Abu-Hanna A, Cornet R, de Keizer NF. Time Spent on Dedicated Patient Care and Documentation Tasks Before and After the Introduction of a Structured and Standardized Electronic Health Record. Appl Clin Inform. 2018 Jan;9(1):46–53. doi:10.1055/s-0037-1615747 PubMed PMID: 29342479; PubMed Central PMCID: PMC5801881.

READY - MADE CITATION

Avinash Gutte, Dixit Varma, Snehal Kose, Abin Jose. LLM-Standardized CTA Impressions for PAD Revascularization Planning: Blinded Single-Center Agreement, Accuracy, And Efficiency Study, *Hell J Radiol 2026*; 11(2): 50-57.