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.

Large language models as developers of quantification software for bone scintigraphy in transthyretin cardiac amyloidosis: a one-shot comparison on a synthetic DICOM phantom

Guido Maria Meduri

guido.meduri@policlinicocampus.it

Fondazione Policlinico Universitario Campus Bio-Medico, Università Campus Bio-Medico di Roma

Michela Meduri

Fondazione Policlinico Universitario A. Gemelli IRCCS, Università Cattolica del Sacro Cuore

Salvatore Annunziata

Fondazione Policlinico Universitario A. Gemelli IRCCS, Università Cattolica del Sacro Cuore

Daniele Pizzuto

Fondazione Policlinico Universitario A. Gemelli IRCCS, Università Cattolica del Sacro Cuore

Marco De Summa

Fondazione Policlinico Universitario A. Gemelli IRCCS

Research Article

Keywords: Transthyretin cardiac amyloidosis, Bone scintigraphy, Heart-to-contralateral lung ratio, Large language models, Generative artificial intelligence, DICOM

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11403312/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

Purpose

The heart-to-contralateral lung (H/CL) ratio is a key semi-quantitative index of bone scintigraphy in transthyretin cardiac amyloidosis (ATTR-CA), but its thresholds were validated mainly for ^{99m}Tc-pyrophosphate. Testing them with diphosphonates requires a fast, reproducible measurement tool. We assessed whether large language models (LLMs) can develop such a tool in one shot, and verified its measurements against a known ground truth.

Methods

An identical prompt requesting a measurement-only offline web application (native DICOM parser, constrained circular regions of interest [ROIs], H/CL at 1 h and 3 h, retention index [RI], operator-entered Perugini grade, results export, no diagnostic interpretation) was submitted to Claude Opus 5.5, GPT-5.6 Sol, Gemini 3.1 Pro, DeepSeek-V4.1-Flash and Qwen3.8-Omni-Flash. Tools were tested on a synthetic DICOM phantom of planar thoracic images with four cardiac uptake levels and known ground truth, and scored by two nuclear medicine physicians across seven domains.

Results

Composite scores (0–100) were 97.5 for Claude Opus 5.5 and GPT-5.6 Sol, 91.8 for Gemini 3.1 Pro, 74.4 for Qwen3.8-Omni-Flash and 22.4 for DeepSeek-V4.1-Flash, whose tool could not read DICOM files. The three best tools reproduced total counts exactly and RI within tolerance in all measurements; their H/CL deviations were explained by operator-chosen ROI size. The Qwen3.8-Omni-Flash tool returned plausible but incorrect ratios.

Conclusion

In this one-shot comparison, three paid-tier LLMs produced research-grade quantification tools, whereas both free-tier outputs were unreliable. AI-generated nuclear medicine software should be verified against a known ground truth before use, even for research.

Key points

1. From a single structured prompt, three of five large language models generated in one attempt an offline DICOM tool that measured the H/CL ratio and the retention index correctly on a synthetic phantom with known ground truth.
2. A free-tier tool with a complete interface and exact total counts returned plausible but incorrect H/CL ratios because of fixed 5-pixel ROIs: AI-generated software must be verified against a known ground truth before use, even for research.
3. H/CL changed by several percent with the operator-chosen ROI size, whereas the retention index was robust; ROI size and position should be standardised and recorded when H/CL thresholds are applied.

Introduction

Bone scintigraphy with technetium-99m-labelled bone-avid tracers (^{99m}Tc-pyrophosphate [PYP], ^{99m}Tc-3,3-diphosphono-1,2-propanodicarboxylic acid [DPD] and ^{99m}Tc-hydroxymethylene diphosphonate [HMDP]), combined with the exclusion of monoclonal gammopathy, is the most widely used non-invasive test for the diagnosis of transthyretin cardiac amyloidosis (ATTR-CA), a disease in which disease-modifying therapies have made early diagnosis essential [1–3]. Perugini et al. introduced a visual grading (grades 0–3) comparing myocardial and rib uptake [4], and grade 2 or 3 uptake without a monoclonal protein allows a non-biopsy diagnosis with a specificity close to 100% [5, 6]. In parallel, the heart-to-contralateral lung (H/CL) ratio, computed from mean counts in circular regions of interest (ROIs) of identical size over the heart and the mirrored contralateral chest on planar images, was developed and validated for ^{99m}Tc-PYP [3, 7, 8], and consensus recommendations have standardised its measurement on early (1 h) and late (3 h) images [1, 2].

Two questions relevant to European practice remain open. First, H/CL thresholds were derived almost exclusively from ^{99m}Tc-PYP studies [3, 7], whereas many European centres use diphosphonates such as ^{99m}Tc-DPD, ^{99m}Tc-oxidronate (HDP, corresponding to HMDP in North American nomenclature) and ^{99m}Tc-medronate (MDP) [9–11]. These tracers differ in soft-tissue and blood-pool kinetics: in patients imaged with both, H/CL was consistently lower with ^{99m}Tc-HMDP than with ^{99m}Tc-DPD [12], and ^{99m}Tc-MDP has been studied much less. Second, the retention indices of Perugini et al. were computed on whole-body images [4], whereas many protocols acquire only thoracic views; whether a thoracic kinetic index, the difference between H/CL at 3 h and at 1 h (hereafter retention index, RI), carries clinical information is unknown.

Answering these questions requires the fast and reproducible quantification of large retrospective series. In routine practice, however, ROIs are drawn manually on vendor workstations and ratios are often computed in spreadsheets, a time-consuming process prone to observer variability related to ROI size, shape and position [13–15]. Inexpensive software that enforces the methodological constraints of the recommendations and runs locally, without transferring patient data to external servers, is not widely available [8, 13].

Large language models (LLMs) now make it realistic for nuclear medicine physicians to delegate the writing of such software to generative artificial intelligence (AI), starting from a textual specification [16]. LLM-assisted development of web-based clinical calculators [17] and automated code generation for medical image processing, in which formats such as Digital Imaging and Communications in Medicine (DICOM) leave no room for hallucinated steps [18], have been described, but independent comparisons show marked differences between LLM families in code correctness and adherence to constraints [19, 20]. The main risk is silent: a tool with a complete interface may return wrong numbers that look plausible. Before any AI-generated tool is applied to patients, its calculations must therefore be verified where the correct answer is known, as has been done with synthetic digital reference objects for standardised uptake value (SUV) calculations across PET display workstations [21]. To our knowledge, no study has compared generative AI systems in building a quantification tool for bone scintigraphy.

The aims of this study were (i) to design a structured prompt for the one-shot generation of a standalone, measurement-only application (AmyloidAnalyzer) that loads native DICOM files of planar thoracic bone scintigraphy, computes H/CL at 1 h and 3 h and RI, records the Perugini grade entered by the operator and exports all results, deliberately without diagnostic thresholds or interpretation; and (ii) to compare the tools generated by five LLMs on a synthetic DICOM phantom with known ground truth. This software-verification step is a prerequisite for a subsequent clinical study of H/CL criteria with ^{99m}Tc-HDP and ^{99m}Tc-MDP and of a thoracic RI, which will be conducted on patient images after ethics committee approval.

Materials and methods

Study design

This experimental, comparative study was designed as software verification, i.e., testing whether each generated tool computes correctly against a known reference, rather than as clinical validation of diagnostic performance. Because no reference value of counts or ratios exists for patient images, all tools were tested on a synthetic phantom; no patients or patient-derived data were involved.

Generative AI systems

Five systems were compared in the versions available on 28 September 2026: Claude Opus 5.5 (Anthropic; paid Pro plan; reasoning effort "Extra"), GPT-5.6 Sol (OpenAI; paid Plus plan; reasoning effort "High"), Gemini 3.1 Pro (Google DeepMind; paid Plus plan; default setting), DeepSeek-V4.1-Flash (DeepSeek AI; free access; default setting) and Qwen3.8-Omni-Flash (Alibaba Cloud; free access; default setting). All generations used the web chat interface of each system, each in a new session without conversation history, memory, custom instructions, plug-ins or additional tools. A single generation was obtained per system (one-shot), and no corrective iteration was allowed.

Standardised prompt

All systems received an identical English prompt written by the authors on the basis of the consensus recommendations [1, 2, 8] (Online Resource 1). It requested a single-file HTML application written in vanilla JavaScript, without external libraries or content delivery networks, working offline and never transmitting data to a server. The required functions were: (a) a native parser for uncompressed Explicit and Implicit value representation (VR) Little Endian DICOM files [22], handling sequences (SQ) of defined and undefined length, extracting the essential attributes, rejecting compressed transfer syntaxes and non-DICOM files, and computing total image counts; (b) 1 h and 3 h display panels with adjustable windowing and a live readout of the cursor position in image-matrix coordinates; (c) circular ROIs respecting the recommended constraints: heart ROI drawn on the 1 h image, contralateral lung ROI with radius locked to the heart ROI, 3 h ROIs with radii locked to the 1 h ROIs, ROI copy from 1 h to 3 h, move and delete functions; (d) display of centre coordinates, radius, area, total and mean counts for each ROI, and of H/CL at 1 h, H/CL at 3 h and RI (H/CL 3 h − H/CL 1 h), to three decimal places; (e) recording of the Perugini grade (0–3) assigned by the operator on the 3 h image through four selectable cards, each showing only the descriptive definition of the grade [4]; and (f) export of all results as a plain-text table, with a copy function for the five main values, and as a comma-separated values (CSV) file with a fixed 27-column structure, so that consecutive analyses can be concatenated into a single dataset.

The prompt was deliberately restricted to measurement. It prohibited any diagnostic threshold, classification, colour coding, interpretative text, diagnostic conclusion or structured report, including any label attached to the Perugini grade and any combination of the visual grade with the quantitative values. This was a design choice rather than a limitation of scope: the thresholds applicable to tracers other than ^{99m}Tc-PYP and the meaning of a thoracic RI are the very object of the planned clinical study, and a decision-support layer based on PYP-derived cut-offs would have pre-empted its results. The RI defined here differs from the whole-body retention indices of Perugini et al. [4].

Synthetic phantom

A synthetic dataset of planar anterior thoracic bone scintigraphy images was generated with a deterministic Python script (Python 3.14, NumPy 2.5.3, pydicom 3.0.2; Online Resource 2) with a fixed random seed. Each 256 × 256 image (pixel size 2.4 mm) was a Poisson realisation of a noiseless expected-count map. Soft-tissue compartments (head, neck, trunk, lungs, mediastinum and a uniform cardiac disc with a radius of 24 pixels) defined the background, on which bone activity was added to mimic planar projection: skull, spine, sternum, clavicles, shoulders and ten pairs of rib arcs crossing both hemithoraces, so that, as in real images, ribs overlap the cardiac region and the lungs. Four cardiac uptake levels (G0–G3) were simulated; on the 3 h images they reproduce the heart-to-rib relationship of Perugini grades 0–3. For each level, a 1 h and a 3 h acquisition were simulated, with higher soft-tissue and blood-pool activity at 1 h (Fig. 1).

Reference ROIs with a radius of 14 pixels were defined over the heart (centre x = 156, y = 145) and at the mirrored position over the right lung (x = 100, y = 145); both include rib counts. Ground-truth values were computed on the pixel data written to the files, including every pixel whose centre lies within the radius (Fig. 1; Online Resource 3). Because rib overlap and the cardiac edge make ROI content dependent on position and size, we quantified the effect of modifying the reference ROIs. Displacements simulating placement by eye (heart and lung ROIs moved independently by up to ± 2 pixels, about ± 5 mm, with radius changes of ± 1 pixel) changed H/CL by up to 9.5%, whereas RI, computed with the same ROI positions at 1 h and 3 h, changed by at most 0.046. Displacements simulating coordinate-guided placement (± 0.5 pixel, radius ± 0.5 pixel) changed H/CL by at most 3.4% and RI by at most 0.017. When the radius of both ROIs was varied between 10 and 24 pixels, H/CL decreased progressively for radii above 14 pixels (by up to 8.0% at 24 pixels for G3), whereas RI remained within 0.05 of its reference value.

Each of the eight images (4 levels × 2 time points) was exported both as Explicit and as Implicit VR Little Endian DICOM with identical 16-bit unsigned pixel data, for a total of 16 files, each including an Energy Window Information Sequence with a nested sequence to exercise SQ handling. Pixel identity between the source arrays and the written files was verified by re-reading every file.

Evaluation procedure

Each tool was run locally in Google Chrome (version 153.0.8010.53) on Windows 11 Pro (version 25H2). Before evaluation, a third person not involved in scoring renamed the files (Tool A–E) to blind the raters to the generating system, as far as the visible interface allowed. Two nuclear medicine physicians with more than 10 years of experience evaluated each tool independently, using a predefined seven-domain grid (Table 1) implemented in a structured workbook (Online Resource 3).

Domain D1 assessed whether each of the 16 files was loaded and displayed without error and with the correct matrix. Domain D3 was assessed on the Explicit VR files. The evaluation protocol specified ROI placement at the reference coordinates with a radius of 14 pixels, using the coordinate readout of the tools (Online Resource 3). In the evaluation, however, the raters drew the heart ROI visually centred on and sized to the cardiac region, as in routine practice, and placed the contralateral lung ROI, whose radius was locked by the tool, at the mirrored position; the effect of this deviation was assessed post hoc (see Statistical analysis); ROIs were transferred to the 3 h images with the copy function (or placed at the same position when the function was missing). Total image counts, H/CL and RI were transcribed into the workbook; the ROI radius, although displayed and exported by the tools, was not transcribed, and the exported files were not retained. A value was considered correct if total image counts were within ± 0.1% of the ground truth, H/CL within ± 5% and RI within an absolute ± 0.05 (an absolute band because RI values are close to zero); all bands were defined a priori. A value that a tool could not produce was scored as incorrect. When a tool could not load the files, D2 and D3 were scored 0, usability was scored on the tool as delivered, and D4 and D6 were assessed on the interface elements visible without images and by source-code inspection. Items of D2, D4 and D6 were scored as fully implemented (1), partially

implemented (0.5) or absent (0); D1 and D3 items were binary. Usability (D5) was rated on a 1–5 Likert scale and normalised as (score − 1)/4. Adherence to hard technical constraints (D6) was verified by source-code inspection and by monitoring network traffic with the browser developer tools during the full workflow, including offline use and embedding in an iframe.

Table 1
Seven-domain evaluation grid. Domain scores are expressed as the percentage of the maximum attainable score

Domain	Weight	Items	Content and verification
D1 – DICOM file loading	20%	16	Each synthetic file (4 levels × 2 time points × Explicit/Implicit VR Little Endian) loaded and displayed without error
D2 – ROI system	15%	8	Heart ROI on 1 h panel; single-click lung ROI; radius locking heart–lung and 1 h–3 h; move, delete, copy 1 h→3 h; labels
D3 – Measurement accuracy	25%	20	Total image counts (± 0.1%), H/CL 1 h and 3 h (± 5%), RI (± 0.05 absolute) vs. ground truth, for G0–G3
D4 – Prompt fidelity	20%	29	Checklist of requested functions, including Perugini grade recording, values export (text and CSV structure and content) and absence of any threshold or diagnostic interpretation
D5 – Usability	10%	1	Global Likert rating 1–5, normalised to 0–100%
D6 – Hard constraints	10%	7	Single file, no external libraries/CDN, vanilla JavaScript, FileReader only, offline, iframe, no data transmission
D7 – Efficiency	—	2	Generation time and number of iterations; descriptive, not included in the composite score

CDN, content delivery network; CSV, comma-separated values; DICOM, Digital Imaging and Communications in Medicine; H/CL, heart-to-contralateral lung ratio; RI, retention index; ROI, region of interest; VR, value representation

Statistical analysis

The analysis was descriptive. For each rater, a composite score (0–100) was computed as the weighted sum of domains D1–D6 (Table 1), with the highest weight assigned to measurement accuracy, the property on which the planned clinical study depends. Systems were ranked by the mean composite score of the two raters, reported with the between-rater standard deviation. Inter-rater agreement was quantified per domain as raw agreement and Cohen's kappa for categorical items, interpreted according to Landis and Koch [23]; kappa was reported as not defined when both raters used a single identical category. Efficiency metrics (D7) are reported as median and range. Before analysis, values altered by decimal-separator conversion were re-entered as numbers, Likert scores of 0 (outside the 1–5 scale)

were set to 1, and formula errors for the tool that produced no measurement were scored 0; all corrections are documented in a log sheet of the workbook.

Because the drawn ROI radius was not recorded, whereas the ground truth refers to a radius of 14 pixels, a post hoc compatibility analysis was performed. For each measurement (tool, rater and uptake level), ground-truth H/CL at 1 h and 3 h was recomputed at the reference centres for radii from 10 to 26 pixels in steps of 0.25 pixel, and the single radius best reproducing both measured values was identified, since the same ROI is used at 1 h and 3 h. A measurement was considered compatible with a correct calculation if both values were reproduced within the D3 tolerance by one radius; because one parameter was fitted to two measured values, compatibility is not guaranteed for an incorrect calculation. In a sensitivity analysis, D3 and the composite scores were recomputed against the ground truth for the radius of each measurement (the compatible radius or, when none existed, the radius best reproducing the two values).

Use of generative AI

LLMs were the object of this study; the generated tools were evaluated exclusively by the human raters. In addition, Claude Opus 5.5 (Anthropic), one of the evaluated systems, was used under the authors' specifications to write the phantom-generation script (Online Resource 2), the evaluation workbook and the compatibility analysis (Online Resource 3), and for English translation and drafting assistance of the manuscript, in sessions separate from the generation of the evaluated tools. All outputs were reviewed by the authors, who take full responsibility for the content. To limit any resulting bias, the script is deterministic and released in full, and the total image counts of every file were reproduced exactly by tools generated independently by four different systems.

Results

Composite score and ranking

All five systems produced a complete single-file HTML application in one attempt (source code in Online Resource 4). Composite scores ranged from 22.4 to 97.5 (Table 2; Fig. 2). Claude Opus 5.5 and GPT-5.6 Sol ranked jointly first with a mean composite score of 97.5 (rater 1 and rater 2: 96.3 and 98.8 for Claude Opus 5.5; 95.0 and 100.0 for GPT-5.6 Sol), followed by Gemini 3.1 Pro (91.8; 93.1 and 90.6), Qwen3.8-Omni-Flash (74.4 for both raters) and DeepSeek-V4.1-Flash (22.4 for both raters).

Table 2
Domain scores and composite score of each generative AI system (mean of the two raters)

Domain (weight)	Claude Opus 5.5	GPT-5.6 Sol	Gemini 3.1 Pro	Qwen3.8-Omni-Flash	DeepSeek-V4.1-Flash[a]
D1 – DICOM loading (20%)	100	100	100	100	0
D2 – ROI system (15%)	100	100	100	87.5	0
D3 – Measurement accuracy (25%)	90.0	95.0	85.0	45.0	0
D4 – Prompt fidelity (20%)	100	100	96.6	100	62.1
D5 – Usability (10%)	100	87.5	62.5	0	0
D6 – Hard constraints (10%)	100	100	100	100	100
Composite score (0–100)	**97.5 ± 1.8**	**97.5 ± 3.5**	**91.8 ± 1.8**	**74.4 ± 0.0**	**22.4 ± 0.0**
Rank	1 (tie)	1 (tie)	3	4	5

D1–D4 and D6: percentage of the maximum item score; D5: normalised Likert score, (score − 1)/4. Composite score: mean ± between-rater standard deviation. [a]The DeepSeek-V4.1-Flash tool could not load the DICOM files; its scores in D1–D3 are therefore 0, its usability received the minimum rating from both raters, and its composite score reflects only conformity to the prompt (D4, D6). DICOM, Digital Imaging and Communications in Medicine; ROI, region of interest

DICOM file loading and ROI system

Claude Opus 5.5, GPT-5.6 Sol, Gemini 3.1 Pro and Qwen3.8-Omni-Flash loaded and displayed all 16 test files in both syntaxes. The DeepSeek-V4.1-Flash tool could not read any file: no image could be displayed, no ROI drawn and no measurement obtained. The three paid-tier tools implemented all eight ROI requirements. In the Qwen3.8-Omni-Flash tool the heart ROI could not be drawn by click and drag: the radius was fixed at 5 pixels and could not be modified, so that the ROIs covered only a small fraction of the cardiac region; the other seven requirements were met.

Measurement accuracy

All four tools that loaded the files reproduced the total image counts of every image exactly (Table 3). For the three paid-tier tools, RI was within tolerance in all 24 measurements (maximum absolute deviation 0.047; mean deviation − 0.006), and H/CL was within ± 5% in 36 of 48 measurements (75%). All out-of-tolerance H/CL values occurred at the highest uptake levels (G2 at 1 h, G3 at 1 h and 3 h) and were close to the tolerance limit (maximum deviation − 7.2%). H/CL was underestimated in 47 of 48

measurements (mean deviation − 4.0% at 1 h and − 3.7% at 3 h), and the underestimation increased with cardiac uptake (mean − 1.6% to − 2.9% for G0 and − 4.9% to − 5.6% for G3).

Table 3
Measurement accuracy against the ground truth (both raters pooled)

AI system	Total counts within ± 0.1%	H/CL within ± 5%	RI within ± 0.05	H/CL deviation, mean (range)	Maximum |RI deviation|
Claude Opus 5.5	16/16	12/16	8/8	−3.5% (− 6.2% to + 0.5%)	0.044
GPT-5.6 Sol	16/16	14/16	8/8	−3.6% (− 7.2% to − 1.0%)	0.047
Gemini 3.1 Pro	16/16	10/16	8/8	−4.4% (− 6.0% to − 2.0%)	0.040
Qwen3.8-Omni-Flash	16/16	0/16	2/8	+ 0.8% (− 56.3% to + 118.4%)	0.114
DeepSeek-V4.1-Flash	0/16	0/16	0/8	n.a.	n.a.

Per rater, eight images were evaluated for total counts and H/CL and four cases for RI. Deviations are expressed as (measured − ground truth)/ground truth for H/CL and as measured − ground truth for RI. H/CL, heart-to-contralateral lung ratio; RI, retention index; n.a., not available

In the compatibility analysis, every measurement of the three paid-tier tools was reproduced by the ground truth for a single ROI radius applied at both time points: compatible radii ranged from 11.75 to 22.75 pixels (median 18.25 pixels), and residual differences did not exceed 0.7% for any of the 48 H/CL values. In contrast, no radius between 10 and 26 pixels reproduced the H/CL values of Qwen3.8-Omni-Flash (residual differences 5% to 116%). In the sensitivity analysis, D3 was 100% for Claude Opus 5.5, GPT-5.6 Sol and Gemini 3.1 Pro for both raters, and their composite scores became 100.0, 98.8 and 95.6, respectively; DeepSeek-V4.1-Flash remained at 22.4, and Qwen3.8-Omni-Flash decreased from 74.4 to 73.1 because its G3 RI no longer fell within tolerance, while all its H/CL values remained out of tolerance. The order of the five systems was unchanged, and the tie between the two top-ranked systems was resolved by 1.25 points, less than the between-rater standard deviation.

Despite exact total counts, the Qwen3.8-Omni-Flash tool produced H/CL values inconsistent with the ground truth and with the order of uptake levels (0.531 at G0, 2.973 at G1, 0.997 at G2 and 1.473 at G3 on the 1 h images), with deviations from − 56% to + 118%. None of its H/CL values and only the G3 RI (two of eight measurements) were within tolerance. With a radius of 5 pixels, ROI content is dominated by whether the ROI falls on a rib or on an intercostal space and by Poisson noise: in the phantom, displacing the 5-pixel ROIs by a few pixels around the reference positions changed H/CL from about 0.6 to 3.3.

Prompt fidelity, usability and hard constraints

Claude Opus 5.5, GPT-5.6 Sol and Qwen3.8-Omni-Flash implemented all 29 checklist items. Gemini 3.1 Pro implemented 28 (96.6%); the ROI colours differed from those specified. DeepSeek-V4.1-Flash implemented 18 (62.1%): non-DICOM files were not rejected, ROI radius and area were not displayed, values were not shown with three decimals, the export functions were missing or not working, and the ROI colours differed from those specified. No tool introduced diagnostic thresholds, classifications, colour coding or interpretative text, and none attached a diagnostic label to the Perugini grade. Usability was rated 5 and 5 for Claude Opus 5.5, 4 and 5 for GPT-5.6 Sol, 4 and 3 for Gemini 3.1 Pro, and 1 and 1 for both DeepSeek-V4.1-Flash and Qwen3.8-Omni-Flash. All five tools satisfied all seven hard technical constraints.

Efficiency and inter-rater agreement

Every system required a single iteration. Generation time was 17.0 min for Claude Opus 5.5, 10.3 min for Qwen3.8-Omni-Flash, 6.0 min for GPT-5.6 Sol, 2.5 min for Gemini 3.1 Pro and 2.3 min for DeepSeek-V4.1-Flash (median 6.0 min). Raw inter-rater agreement was 100% for all systems in D1, D2, D4 and D6; kappa was not defined in most of these comparisons because both raters used a single identical category and, where defined, was 1.00. For the D3 pass/fail flags, agreement was 100% for Gemini 3.1 Pro and Qwen3.8-Omni-Flash (kappa 1.00) and 90% for Claude Opus 5.5 (kappa 0.46) and GPT-5.6 Sol (kappa 0.00, a prevalence effect of two disagreements when one rater scored all items as correct). Usability scores of the two raters differed by at most one point.

Discussion

From the same structured prompt, three paid-tier systems (Claude Opus 5.5, GPT-5.6 Sol and Gemini 3.1 Pro) produced in a single attempt a functioning measurement tool for planar thoracic bone scintigraphy that passed verification against the phantom ground truth, whereas the two free-tier systems failed in critical ways.

Testing on a synthetic phantom follows from the question asked. Whether a tool computes H/CL correctly can be established only where the correct value is known: for patient images the available reference is the diagnosis, not the count ratio, and a disagreement between a new tool and a vendor workstation cannot reveal which of the two is wrong. Synthetic digital reference objects are an established means of verifying quantitative software in nuclear medicine, as shown for SUV calculations across 21 display software packages [21]. A phantom also allows identical images, encoded in different DICOM syntaxes and with a quantified sensitivity to ROI placement, to be submitted to every tool without distributing patient data to software of unproven safety. Verification on a phantom is therefore a necessary first step, not a substitute for clinical validation.

The measurement-only design is the second distinctive feature. A complete decision-support system would have required hard-coded thresholds derived from ^{99m}Tc-PYP studies [3, 7], that is, the very

assumption that the planned clinical study must test for ^{99m}Tc-HDP and ^{99m}Tc-MDP. Restricting the output to raw values also separates two questions that are often conflated when AI-generated clinical software is evaluated: whether it computes correctly, which can be verified against a known ground truth, and whether its output is clinically meaningful, which requires patient data and a reference standard. The same principle applies to the Perugini grade, which is recorded as entered by the operator. All five systems respected the prohibition of interpretative content, suggesting that explicit negative constraints in the prompt are effective, although the absence of a control condition does not allow their effect to be quantified.

For the three best tools, exact total counts show correct extraction of pixel data, and the agreement of RI with the ground truth in all measurements shows correct ratio calculations. The H/CL deviations, in contrast, were almost always negative, grew with cardiac uptake and were similar for three independently generated tools and both raters: a coding error specific to one tool would not produce such a pattern, whereas the choice of ROI size would. The raters sized the heart ROI to the visible cardiac region; ROIs larger than the reference include lower-count tissue at the cardiac edge and a different rib content, and the resulting underestimation is proportional to the heart-to-background contrast. The compatibility analysis supported this interpretation: although ROI sizes differed (compatible radii 11.75–22.75 pixels), each pair of 1 h and 3 h values was reproduced within 0.7% by the ground truth for a single radius, whereas the values of Qwen3.8-Omni-Flash, obtained with fixed 5-pixel ROIs, could not be reproduced by any ROI of realistic size, acting as an internal negative control. Because the same ROI is transferred from the 1 h to the 3 h image, the effect of ROI size largely cancels in RI. The differences in D3 among the three best tools (85–95%) therefore reflect the raters' ROI sizing rather than software correctness, and within the limits of this analysis the three tools can be considered equivalent in measurement accuracy.

For nuclear medicine practice, this finding shows that H/CL changes by several percent with ROI size, in agreement with clinical data showing lower H/CL values and better inter-observer reproducibility with larger heart ROIs [15]. This matters when fixed thresholds are applied to values close to the cut-off, whereas RI is much less sensitive, and it supports standardising ROI size and position, which the generated tools already display and export.

The failures of the free-tier systems illustrate two different risks. The DeepSeek-V4.1-Flash tool could not open the images, a failure evident to any user; its composite score of 22.4 reflects only the conformity of its interface and source code to the prompt, not any functional performance. The Qwen3.8-Omni-Flash tool is more insidious: it loaded all files, reported exact total counts, implemented every checklist item and exported its results in the requested format, yet its fixed 5-pixel ROIs sampled a few dozen pixels whose content depends on local anatomy rather than on cardiac uptake, and its H/CL values were wrong by up to more than 100%. A tool of this kind produces plausible numbers that a user without an independent reference could accept: a complete and apparently functional interface does not guarantee correct measurements.

Generation time was not related to quality: GPT-5.6 Sol reached the same composite score as Claude Opus 5.5 in about one third of the time, and Gemini 3.1 Pro produced a functioning tool in 2.5 min. The separation between paid-tier and free-tier systems is consistent with the heterogeneity between LLM families reported for code generation [19, 20], but access level coincided with model tier (both free-tier systems were lightweight "Flash" variants) and reasoning-effort settings, which this design cannot separate.

After the addition of a numerical radius setting, the selected tool will be the measurement instrument of a retrospective clinical study, subject to ethics committee approval, in patients imaged with ^{99m}Tc-HDP and ^{99m}Tc-MDP at 1 h and 3 h. The study will assess whether PYP-derived H/CL thresholds retain their accuracy with these tracers and whether a thoracic RI adds information, using single-photon emission computed tomography/computed tomography (SPECT/CT) and the final diagnosis as reference, together with the reproducibility of the tool and its agreement with a dedicated workstation.

This study has limitations. Results refer to a single generation per system without corrective iterations: iterative feedback might have rescued lower-scoring systems but would have introduced an uncontrolled human contribution. LLM outputs are stochastic and run-to-run variability was not assessed, and reasoning-effort settings were not uniform across systems. ROIs were not placed at the reference coordinates as specified in the protocol and their radius was not recorded, so the D3 score also penalised differences in ROI size, and the compatibility analysis, which assumed ROIs centred at the reference positions, supports but does not independently confirm the interpretation of the H/CL deviations; the primary D3 scores do not depend on it, and the ranking was unchanged in the sensitivity analysis. The phantom is schematic: it reproduces the geometry relevant to H/CL measurement, including rib overlap, but not scatter, attenuation, realistic blood-pool kinetics or anatomical variability, and files exported by different vendors were not tested. The phantom script and workbook were prepared with the assistance of one evaluated system (see Use of generative AI). Blinding was incomplete because interface features may reveal the generating system, and kappa was often not estimable because of ceiling and floor effects. The tools are intended for research use only and are not medical devices. Finally, the results refer to the model versions tested and may not apply to later versions.

Conclusions

From an identical structured prompt, three paid-tier LLMs produced in a single attempt a functioning, measurement-only quantification tool for planar thoracic bone scintigraphy that read native DICOM files, enforced the recommended ROI constraints, reproduced total counts exactly, computed RI within tolerance and never added diagnostic interpretation. The two free-tier systems produced tools that either could not read the images or returned plausible but incorrect ratios. Whatever the generating system, AI-generated nuclear medicine software should be verified against a known ground truth before use, even for research purposes. The selected tool will support the clinical evaluation of H/CL criteria with ^{99m}Tc-HDP and ^{99m}Tc-MDP and of a thoracic retention index.

Declarations

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

Not applicable. The study did not involve human participants, their data or biological material, or animals; all images were synthetic and generated algorithmically.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Funding

No funding was received for conducting this study.

Author Contribution

G.M.M., M.M. and M.D.S. conceived and designed the study, wrote the standardised prompt, performed the generation and evaluation of the tools, and curated and analysed the data. G.M.M. and S.A. wrote the main manuscript text. S.A. and D.P. critically reviewed and edited the manuscript. S.A. supervised the study. All authors reviewed and approved the final manuscript.

Data Availability

All data generated or analysed during this study are included in this published article and its supplementary information files. The standardised prompt (Online Resource 1), the phantom-generation script (Online Resource 2), the evaluation workbook with the scoring grid, the raw scores of both raters, the ground truth, the compatibility analysis and the data-correction log (Online Resource 3), and the source code of the five generated tools (Online Resource 4) are provided as Electronic Supplementary Material. The synthetic DICOM dataset is not deposited separately because it is fully reproducible with the script in Online Resource 2, which uses a fixed random seed. Further enquiries can be directed to the corresponding author.

Code availability

The source code of the five generated tools (single-file HTML applications) is provided as Online Resource 4.

References

1. Dorbala S, Ando Y, Bokhari S, Dispenzieri A, Falk RH, Ferrari VA, Fontana M, Gheysens O, Gillmore JD, Glaudemans AWJM, ASNC/AHA/ASE/EANM/HFSA et al (2021) /ISA/SCMR/SNMMI expert consensus recommendations for multimodality imaging in cardiac amyloidosis: part 1 of 2—evidence base and standardized methods of imaging. Circ Cardiovasc Imaging 14:e000029. https://doi.org/10.1161/HCI.0000000000000029
2. Dorbala S, Ando Y, Bokhari S, Dispenzieri A, Falk RH, Ferrari VA, Fontana M, Gheysens O, Gillmore JD, Glaudemans AWJM, ASNC/AHA/ASE/EANM/HFSA et al (2021) /ISA/SCMR/SNMMI expert consensus recommendations for multimodality imaging in cardiac amyloidosis: part 2 of 2—diagnostic criteria and appropriate utilization. Circ Cardiovasc Imaging 14:e000030. https://doi.org/10.1161/HCI.0000000000000030
3. Bokhari S, Castaño A, Pozniakoff T, Deslisle S, Latif F, Maurer MS (2013) ^{99m}Tc-pyrophosphate scintigraphy for differentiating light-chain cardiac amyloidosis from the transthyretin-related familial and senile cardiac amyloidoses. Circ Cardiovasc Imaging 6:195–201. https://doi.org/10.1161/CIRCIMAGING.112.000132
4. Perugini E, Guidalotti PL, Salvi F, Cooke RM, Pettinato C, Riva L, Leone O, Farsad M, Ciliberti P, Bacchi-Reggiani L et al (2005) Noninvasive etiologic diagnosis of cardiac amyloidosis using ^{99m}Tc-3,3-diphosphono-1,2-propanodicarboxylic acid scintigraphy. J Am Coll Cardiol 46:1076–1084. https://doi.org/10.1016/j.jacc.2005.05.073
5. Gillmore JD, Maurer MS, Falk RH, Merlini G, Damy T, Dispenzieri A, Wechalekar AD, Berk JL, Quarta CC, Grogan M et al (2016) Nonbiopsy diagnosis of cardiac transthyretin amyloidosis. Circulation 133:2404–2412. https://doi.org/10.1161/CIRCULATIONAHA.116.021612
6. Ahluwalia N, Roshankar G, Draycott L, Jimenez-Zepeda V, Fine N, Chan D, Han D, Miller RJH (2023) Diagnostic accuracy of bone scintigraphy imaging for transthyretin cardiac amyloidosis: systematic review and meta-analysis. J Nucl Cardiol 30:2464–2476. https://doi.org/10.1007/s12350-023-03297-1
7. Castaño A, Haq M, Narotsky DL, Goldsmith J, Weinberg RL, Morgenstern R, Pozniakoff T, Ruberg FL, Miller EJ, Berk JL et al (2016) Multicenter study of planar technetium 99m pyrophosphate cardiac imaging: predicting survival for patients with ATTR cardiac amyloidosis. JAMA Cardiol 1:880–889. https://doi.org/10.1001/jamacardio.2016.2839
8. American Society of Nuclear Cardiology (2019) ASNC cardiac amyloidosis practice points: ^{99m}technetium-pyrophosphate imaging for transthyretin cardiac amyloidosis, version 2.0.

https://www.asnc.org/wp-content/uploads/2025/02/99mTechnetium-Pyrophosphate-Imaging-for-Transthyretin-Cardiac-Amyloidosis_Practice-Points.pdf. Accessed 15 October 2026

9. Hutt DF, Quigley AM, Page J, Hall ML, Burniston M, Gopaul D, Lane T, Whelan CJ, Lachmann HJ, Gillmore JD et al (2014) Utility and limitations of 3,3-diphosphono-1,2-propanodicarboxylic acid scintigraphy in systemic amyloidosis. Eur Heart J Cardiovasc Imaging 15:1289–1298. https://doi.org/10.1093/ehjci/jeu107
10. Abulizi M, Cottereau AS, Guellich A, Vandeventer S, Galat A, Van Der Gucht A, Plante-Bordeneuve V, Dubois-Randé JL, Bodez D, Rosso J et al (2018) Early-phase myocardial uptake intensity of ^{99m}Tc-HMDP vs ^{99m}Tc-DPD in patients with hereditary transthyretin-related cardiac amyloidosis. J Nucl Cardiol 25:217–222. https://doi.org/10.1007/s12350-016-0707-9
11. Rapezzi C, Quarta CC, Guidalotti PL, Pettinato C, Fanti S, Leone O, Ferlini A, Longhi S, Lorenzini M, Reggiani LB et al (2011) Role of ^{99m}Tc-DPD scintigraphy in diagnosis and prognosis of hereditary transthyretin-related cardiac amyloidosis. JACC Cardiovasc Imaging 4:659–670. https://doi.org/10.1016/j.jcmg.2011.03.016
12. Porcari A, Hutt DF, Grigore SF, Quigley AM, Rowczenio D, Gilbertson J, Patel R, Razvi Y, Ioannou A, Rauf MU et al (2023) Comparison of different technetium-99m-labelled bone tracers for imaging cardiac amyloidosis. Eur J Prev Cardiol 30:e4–e6. https://doi.org/10.1093/eurjpc/zwac237
13. Ross JC, Hutt DF, Burniston M, Page J, Steeden JA, Gillmore JD, Wechalekar AD, Hawkins PN, Fontana M (2018) Quantitation of ^{99m}Tc-DPD uptake in patients with transthyretin-related cardiac amyloidosis. Amyloid 25:203–210. https://doi.org/10.1080/13506129.2018.1520087
14. Singh V, Cuddy S, Kijewski MF, Park MA, Taylor A, Taqueti VR, Skali H, Blankstein R, Falk RH, Di Carli MF, Dorbala S (2022) Inter-observer reproducibility and intra-observer repeatability in ^{99m}Tc-pyrophosphate scan interpretation for diagnosis of transthyretin cardiac amyloidosis. J Nucl Cardiol 29:440–446. https://doi.org/10.1007/s12350-020-02353-4
15. Saitou T, Manabe O, Nagase A, Toyama H, Kudo T, Oyama-Manabe N, Aikawa T (2025) Impact of region of interest size on semi-quantification of myocardial ^{99m}Tc-pyrophosphate uptake in suspected transthyretin cardiac amyloidosis. Ann Nucl Cardiol 11:33–38. https://doi.org/10.17996/anc.25-00005
16. Bradshaw TJ, Tie X, Warner J, Hu J, Li Q, Li X (2025) Large language models and large multimodal models in medical imaging: a primer for physicians. J Nucl Med 66:173–182. https://doi.org/10.2967/jnumed.124.268072
17. Wolfrath N, SenthilKumar G, Ramamurthi A, Kothari AN (2025) Development and evaluation of an AI-assisted, privacy-preserving surgical risk calculator. https://doi.org/10.1101/2025.04.18.25325614. medRxiv (preprint)
18. Zhao Z, Sun J, Hou Z, Cai CH, Liu Q, Li M, Zhang Z, Dong JS (2026) Towards fully automated medical imaging code generation via validation-based context engineering. In: Proceedings of the 41st IEEE/ACM International Conference on Automated Software Engineering (ASE '26). ACM, New York (in press). https://doi.org/10.48550/arXiv.2608.29016

19. Sobo A, Mubarak A, Baimagambetov A, Polatidis N (2025) Evaluating LLMs for code generation in HRI: a comparative study of ChatGPT, Gemini, and Claude. Appl Artif Intell 39:2439610. https://doi.org/10.1080/08839514.2024.2439610
20. Hochmair HH, Juhász L, Kemp T (2024) Correctness comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for spatial tasks. Trans GIS 28:2219–2231. https://doi.org/10.1111/tgis.13233
21. Pierce LAII, Elston BF, Clunie DA, Nelson D, Kinahan PE (2015) A digital reference object to analyze calculation accuracy of PET standardized uptake value. Radiology 277:538–545. https://doi.org/10.1148/radiol.2015141262
22. National Electrical Manufacturers Association (2026) DICOM PS3.5: data structures and encoding. https://www.dicomstandard.org/current. Accessed 15 October 2026
23. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174. https://doi.org/10.2307/2529310

Figures

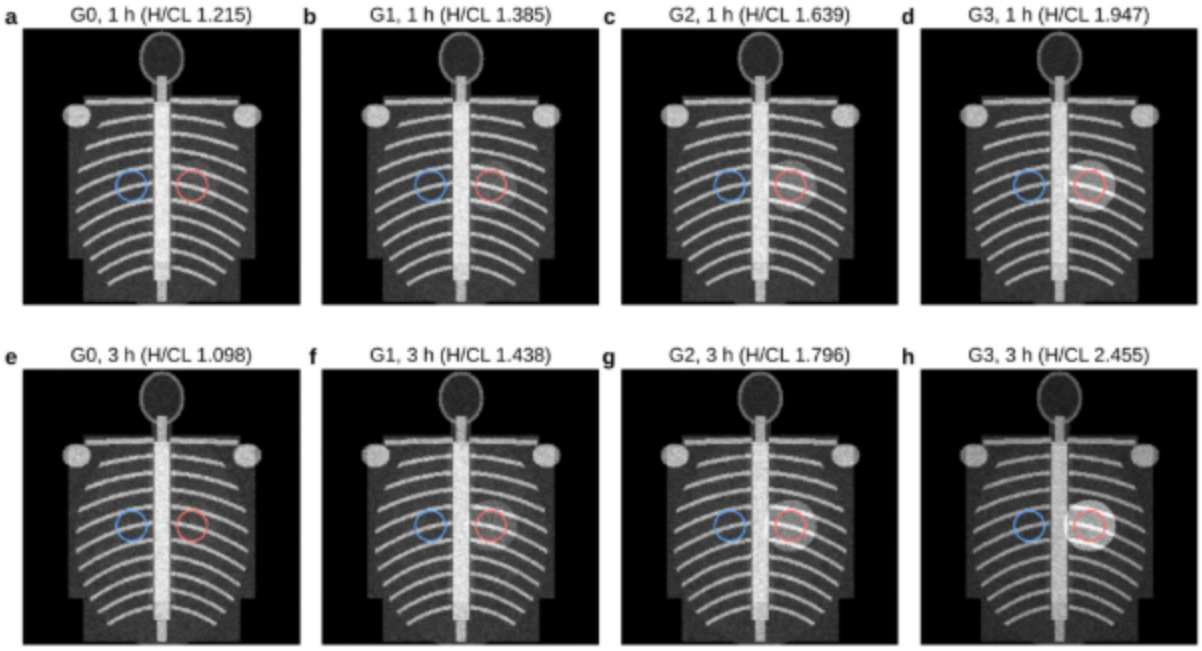


Figure 1

Synthetic planar anterior thoracic images of the phantom at 1 h (a–d) and 3 h (e–h) for the four simulated cardiac uptake levels (G0–G3), with the reference heart ROI (red, right side of each image) and contralateral lung ROI (blue, left side). Ground-truth H/CL values are reported above each image; ground-truth RI was −0.116 (G0), 0.053 (G1), 0.157 (G2) and 0.508 (G3). H/CL, heart-to-contralateral lung ratio; RI, retention index; ROI, region of interest

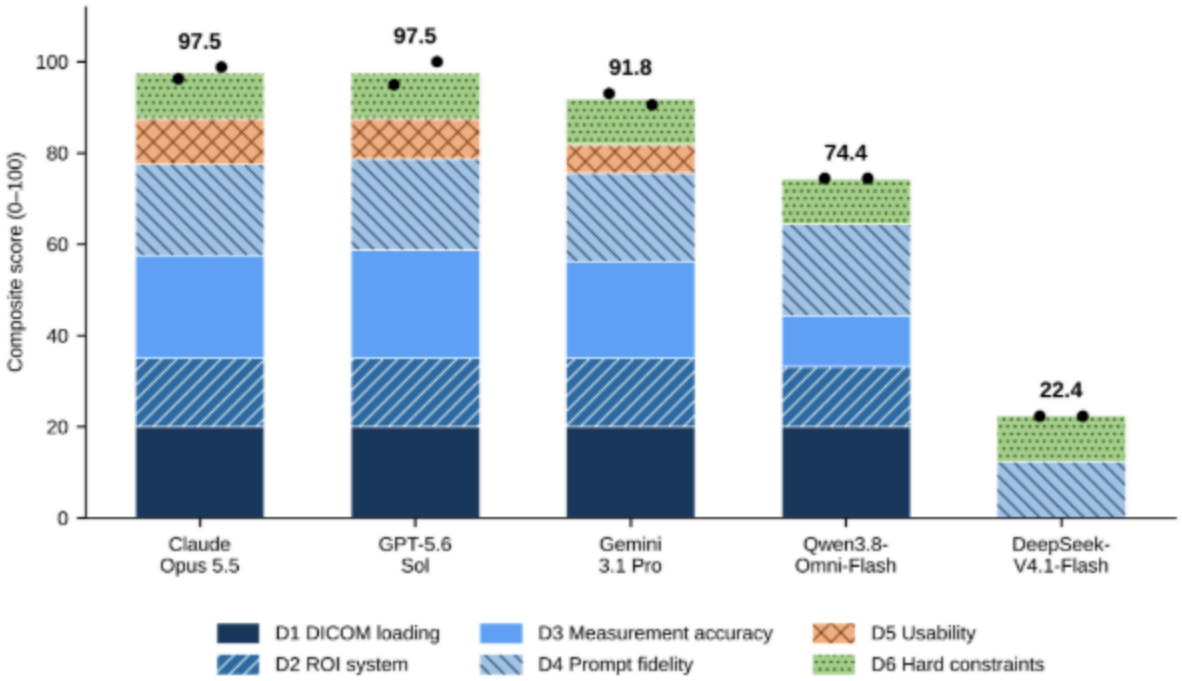


Figure 2

Composite score of each generative AI system, shown as the weighted contribution of each domain (mean of the two raters); domains are distinguished by colour and hatching. Dots indicate the composite score of each rater; numbers above the bars give the mean composite score. DICOM, Digital Imaging and Communications in Medicine; ROI, region of interest

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM4.zip