Open Academic Journal of Advanced Science and Technology
ISSN 2577-7807
Vol. 10, No. 1, pp. 48-59.
2026
DOI: 10.33094/oajast.v10i1.2523
© 2026 by the authors; licensee Online Academic Press, USA

From single slices to whole studies: A scoping review of frontier LLM evaluation on CT and MRI and the case for study-level assessment

Peter Nemeth[1]
Enikő Orsi[2]
Balázs Patczai[3]
Tamás Riedling[4]
Gyula Marada[5]
Zoltán Nyárády[6*]

[1]*ReadYourLab, Győr, Hungary.*
Email: peter@readyourlab.com
[2,4,5]*University of Pécs, Department of Dentistry, Oral and Maxillofacial Surgery, Pécs, Hungary.*
Email: orsi.eniko@pte.hu
Email: riedling.tamas@pte.hu
Email: marada.gyula@pte.hu
[3]*University of Pécs, Department of Traumatology and Hand Surgery, Pécs, Hungary.*
Email: patczai.balazs@pte.hu
[6]*Széchenyi István University, Department of Speech Therapy, Győr, Hungary; ReadYourLab, Győr, Hungary.*
Email: zoltan.nyarady@hotmail.com

Licensed:
This work is licensed under a Creative Commons Attribution 4.0 License.

Keywords:
Benchmark evaluation
Computed tomography
Large language models
Magnetic resonance imaging
Vision–language models.

Received: *27 August 2026*
Revised: *17 September 2026*
Accepted: *20 October 2026*
Published: *23 October 2026*
(Corresponding Author)*

Abstract

Large multimodal and vision–language models are increasingly evaluated on computed tomography (CT) and magnetic resonance imaging (MRI), yet a radiology study is a multi-series, multi-plane, multi-sequence volumetric object read with prior examinations and clinical context, not a single picture. This scoping review maps the 2025–2026 evidence base on frontier and specialist model evaluation for CT and MRI, characterises the granularity at which models are assessed, and examines whether reported scores represent real value for patients and radiologists. Following PRISMA-ScR, we searched arXiv, medRxiv, bioRxiv, PubMed/MEDLINE, IEEE Xplore, major conference proceedings, challenge registries, and dataset repositories, and charted benchmark, dataset, model, and evaluation papers against a four-level input-granularity taxonomy (single slice; single volume; curated multi-series; whole study with priors and context) and realism-gap dimensions. Evaluation clusters at the single-slice and single-volume levels; whole-study evaluation is rare and, where present, relies on degraded imaging with the radiology report available as an input. Frontier models receive volumetric data as sampled slices or montages constrained by per-request image caps and consistently underperform specialist models and radiologists on volumetric and measurement-dependent tasks, while widely reported metrics map weakly onto clinical value and slice-level data leakage can inflate accuracy by tens of percentage points. The deployable-at-scale option, API-accessible frontier models, is therefore the least rigorously evaluated on whole studies. We propose desiderata and a twelve-item checklist for study-level, workflow-faithful evaluation: whole-study input with priors, the report withheld, consensus- or outcome-based reference standards, task framing that mirrors the read, and explicit safety and deployability reporting.

Funding: This research was funded by ReadYourLab Kft., Győr, Hungary.
Institutional Review Board Statement: Not applicable.
Transparency: The authors declare that the manuscript is honest, truthful and transparent, that no important aspects of the study have been omitted and that all deviations from the planned study have been made clear. This study followed all rules of writing ethics.
Competing Interests: P.N. is the founder of ReadYourLab Kft. and Z.N. is affiliated with ReadYourLab Kft., a company developing radiology artificial-intelligence software. To mitigate this competing interest, the review names no company, product, or commercial tool; eligibility criteria and a rejection log were fixed before synthesis; and all recommendations are framed as field-level desiderata derived from the charted evidence. The other authors declare no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Data Availability Statement: The corresponding author may provide study data upon reasonable request.
Use of AI Tools: Drafting and literature charting were assisted by an AI system under author supervision; all included references were verified by the authors against primary sources, and the authors are responsible for the content.

© 2026 by the authors; licensee Online Academic Press, USA

1. Introduction

The clinical promise of multimodal large language models (LLMs) in radiology is now concrete enough to be tested rather than merely asserted. In the space of roughly eighteen months, frontier general-purpose models and purpose-built medical vision–language models (VLMs) have been evaluated on tasks spanning abnormality classification, lesion detection, radiology report generation, visual question answering (VQA), and grounded spatial reasoning across computed tomography (CT) and magnetic resonance imaging (MRI). A 2025–2026 wave of benchmarks and datasets has made these comparisons routine, and a parallel wave of commentary has begun to ask whether the resulting numbers mean what they appear to mean [1].

This review starts from a simple observation about granularity. A radiology *study* — the unit a radiologist is assigned, opens, and signs off — is not a single image. A contrast-enhanced abdominal CT comprises multiple reconstructed series across several contrast phases; a brain MRI comprises several pulse sequences (for example T1, T2, FLAIR, diffusion-weighted imaging with apparent diffusion coefficient maps, and post-contrast T1) in multiple planes; and both are read against the patient's prior imaging and clinical history. The interpretive act involves detecting an abnormality across contiguous slices, localising it, characterising it using the interplay of sequences or phases, comparing it with priors, and synthesising an impression that answers a clinical question. Much of the evaluation literature, by contrast, presents a model with a single exported 2D slice, or at most a single reconstructed volume from one series, and scores its output against a curated label.

This granularity mismatch matters for three reasons. First, it can inflate apparent performance: a positive finding pre-localised to one slice removes the search problem that dominates real reading. Second, it obscures failure modes that only appear when a model must integrate evidence across slices, series, and time. Third, and most importantly, it means the score does not measure the value the model would actually deliver to a radiologist or a patient. The problem is timely because the models most likely to be deployed at scale are the closed, API-accessible frontier systems — cheap to call, requiring no local graphics-processing-unit (GPU) infrastructure — yet these are precisely the models least rigorously evaluated on whole studies.

We therefore ask five questions. RQ1 (granularity): at what level of input granularity are frontier and specialist models currently evaluated on CT/MRI? RQ2 (datasets): what 2025–2026 datasets and benchmarks exist, and how does the input each provides compare with what a radiologist receives for the same clinical question? RQ3 (metrics): what metrics are used, and how well do they represent real value versus curated-label accuracy? RQ4 (frontier vs specialist): how do closed frontier models compare with open specialist 3D models, on measured performance *and* on deployability? RQ5 (the gap and the path): where and why does benchmark performance diverge from real-world radiology, and what would study-level evaluation require?

We address these through a scoping review, using a four-level granularity taxonomy as the analytical spine.

Section 2 sets out what a radiology study is and defines the taxonomy.

Section 3 describes methods.

Sections 4–8 chart datasets, benchmarks, frontier-versus-specialist evidence, metric-to-value mapping, and deployability.

Section 9 synthesises the clinical-realism gap.

Section 10 presents the constructive proposal — desiderata and a checklist for study-level evaluation — followed by limitations and conclusion.

2. Background: What a Radiology Study Actually Is

2.1. Anatomy of a Study

A cross-sectional imaging study is a structured, hierarchical object. At acquisition, a scanner produces one or more *series*; each series is a stack of contiguous cross-sectional images reconstructed at a defined slice thickness and, for CT, a defined kernel and window. A single clinical CT examination commonly contains multiple series that differ in contrast *phase* (non-contrast, arterial, portal-venous, delayed), in reconstruction (soft-tissue, lung, and bone kernels), and in plane (axial, coronal, sagittal reformats). An MRI examination contains multiple *sequences* — for example T1-weighted, T2-weighted, FLAIR, diffusion-weighted imaging (DWI) with apparent-diffusion-coefficient (ADC) maps, and pre- and post-contrast series — each sensitising the image to different tissue properties, again across multiple planes. The full object is encoded in DICOM, with metadata describing geometry, acquisition parameters, and series relationships, and is read alongside *prior* studies for comparison.

Contrast this with the input used in much of the evaluation literature: a single exported PNG of one axial slice, stripped of its neighbours, its companion series, its priors, and its acquisition metadata. The distance between these two objects is the central concern of this review.

2.2. A Taxonomy of Evaluation Granularity

We chart every included item against the level of input granularity actually provided to the model under evaluation (Table 1 uses this axis).

© 2026 by the authors; licensee Online Academic Press, USA

- Level 0 — Single 2D slice. One pre-selected cross-sectional image, often a key slice on which the relevant finding is already visible.
- Level 1 — Single volume / single series. One reconstructed 3D volume from a single series (for example a non-contrast chest CT), presented natively or as sampled slices.
- Level 2 — Multiple series / multiparametric (curated subset). A fixed set of series for one body part and clinical question — for example a four-sequence brain MRI, a biparametric prostate MRI, or a multi-phase liver CT — but still curated to a predetermined series set.
- Level 3 — Whole study / complete exam. All clinically relevant series, phases, and planes for the examination, read with prior studies and clinical context, as a radiologist receives them.

This taxonomy is deliberately about *input*, not model architecture. A 2D model can be evaluated at Level 2 by receiving several series; a 3D model can be evaluated at Level 1 if given only one volume. The taxonomy measures the realism of the evaluation, not the sophistication of the system.

2.3. What "Reading" Means Clinically

Clinically, reading a study is not image classification. It is a sequence of coupled operations: *detection* (finding an abnormality that may occupy only a few of hundreds of slices), *localisation* (assigning it to an anatomical structure and laterality), *characterisation* (using the behaviour of a lesion across sequences or phases to narrow a differential), *comparison* (judging change against priors), and *synthesis* (producing an impression that answers the referring question and triggers management). A benchmark that presents a positive slice and asks "is there a nodule?" removes detection, localisation, comparison, and synthesis in one stroke. The granularity taxonomy is, in effect, a proxy for how many of these operations a benchmark preserves. Figure 1 illustrates this alignment: each step up the taxonomy restores one or more of the reading operations that a lower level removes.

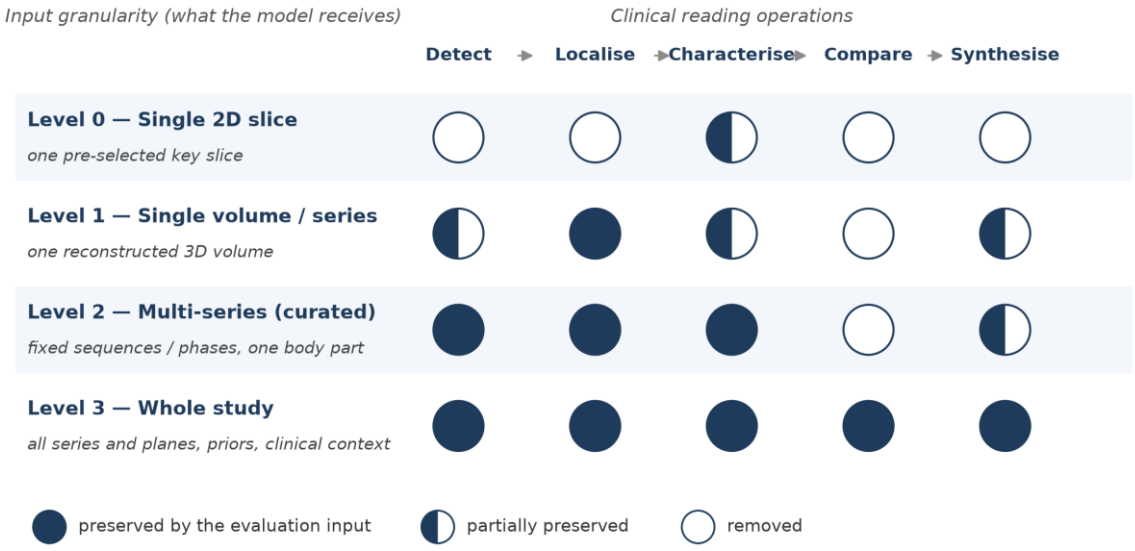


Figure 1. The input-granularity taxonomy (Level 0, single slice, to Level 3, whole study) aligned to the clinical reading operations (detect → localise → characterise → compare → synthesise) that each level preserves (filled), partially preserves (half-filled), or removes (open). Schematic.

3. Methods

3.1. Design and Reporting

We conducted a scoping review with narrative synthesis, following the PRISMA extension for Scoping Reviews (PRISMA-ScR). A scoping design is appropriate because the literature is fast-moving, largely preprint, and heterogeneous in task and metric; the goal is to *map* the field and its gaps rather than to pool effect estimates. The protocol (objectives, eligibility, charting schema) was fixed before synthesis. Because one author has a competing interest (Section: Declarations), eligibility criteria and a rejection log were maintained to protect against selection bias, and no commercial tool or company is named anywhere in the review.

3.2. Eligibility Criteria

Inclusion: (i) Benchmark/dataset papers, model papers reporting evaluation, or evaluation/perspective papers on methodology; (ii) CT and/or MRI, including multiparametric MRI and multi-phase CT; (iii) tasks

© 2026 by the authors; licensee Online Academic Press, USA

of classification, detection, report generation, VQA, retrieval, or grounded reasoning; (iv) published 2025–2026.

Background exception: Pre-2025 works were retained only where a 2025–2026 study builds directly on them (foundational datasets, seminal leakage analyses, established metrics).

Exclusion: Pure segmentation-only papers (unless used to make a granularity point); plain radiography, mammography, or ultrasound unless bundled into a multimodal comparison; non-English; editorials without evaluable content; and single-institution tool papers with no benchmark contribution.

3.3. Information Sources and Search

We searched, in June–July 2026: preprint servers (arXiv — cs.CV, cs.CL, cs.LG, eess.IV; medRxiv; bioRxiv); bibliographic databases (PubMed/MEDLINE, Scopus, IEEE Xplore, Web of Science); conference proceedings (MICCAI, NeurIPS Datasets & Benchmarks, CVPR/ICCV/ECCV medical workshops, ML4H, MIDL); and challenge/registry sources (grand-challenge.org, Hugging Face datasets, Papers-with-Code). Search terms combined four blocks with AND across blocks and OR within: a *model* block (large language model, LLM, vision–language model, VLM, multimodal model, foundation model, and named frontier families), a *modality/object* block (CT, computed tomography, MRI, magnetic resonance, volumetric, 3D medical, whole study, multi-series, multiparametric, multiphase), a *task/evaluation* block (benchmark, evaluation, report generation, VQA, classification, detection, diagnostic accuracy), and a *gap* block (slice-level, volume-level, study-level, data leakage, clinical workflow, real-world). Backward and forward citation chasing was performed on anchor papers.

3.4. Selection and Charting

Records were screened at title/abstract and then full text against the criteria above. For each included item we charted: name, year, venue, and identifier; modality, body region, and task; input-granularity level (0–3); dataset size and label provenance; input representation given to the model (native volume, sampled slices, montage, slice-video, full DICOM stack; single- or multi-series); whether the radiology report was available as an input (report-leakage flag); models evaluated (frontier/specialist) and access mode (API/self-hosted); metrics; comparator (radiologist, ward team, other models, none); and stated realism limitations.

3.5. PRISMA-ScR Flow

The documented search identified records across the sources above; after de-duplication, titles/abstracts were screened, full texts assessed, and a core evidence set of benchmark, dataset, model, and methodology papers was charted, supplemented by a small number of pre-2025 background references retained under the exception in 3.2. Because this is a rapid, preprint-heavy scoping exercise in a field with rapid version churn, the charted set should be read as a snapshot rather than an exhaustive census (Section 11).

4. Datasets for CT/MRI Model Evaluation (2025–2026)

Table 1 charts representative datasets by granularity level. The dominant pattern is Level 1: a single reconstructed volume, most often a non-contrast chest CT, paired with a report or report-derived labels.

Table 1. Representative CT/MRI datasets for model evaluation (2025–2026), charted by input-granularity level.

Dataset	Modality / region	Granularity	Size	Label provenance	Report available as input?	Ref.
CT-RATE (with CT-CLIP / CT-CHAT)	Chest CT (non-contrast)	Level 1	25,692 volumes / 21,304 patients	Paired radiology reports	Yes (paired report)	Hamamci, et al. [2]
RadFM / RadBench	Mixed 2D & 3D radiology	Level 1 (mostly)	13M 2D images; 615K 3D scans	Reports, curated tasks	Varies	Wu, et al. [3]
Merlin	Abdominal CT	Level 1	25,528 CT + EHR + reports	Reports, EHR, codes	Yes (training)	Blankemeier, et al. [4]
Med3DVLM (eval sets)	3D CT	Level 1	Retrieval / VQA / report tasks	Report-derived	Varies	Xin, et al. [5]
MCT-LTDiag	Liver CT (4-phase)	Level 2	517 cases / ~2,068 sequences	Expert tumour labels (5 classes)	No	Wu, et al. [6]

© 2026 by the authors; licensee Online Academic Press, USA

Dataset	Modality / region	Granularity	Size	Label provenance	Report available as input?	Ref.
PI-CAI	Prostate bpMRI	Level 2	1,000-scan hidden test; ~10k dev	Histopathology / expert	No	Saha, et al. [7]
fastMRI+	Knee & brain MRI	Level 2	16,154 + 7,570 box annotations	Subspecialist annotations	No	Zhao, et al. [8]
INSPECT	CT pulmonary angiography	Level 3 (bundled clinical)	23,248 CTPA / 19,402 patients	Report impressions + EHR/outcomes	Yes (impression)	Huang, et al. [9]
VALID (evaluation set)	Multimodal inpatient (CT/MRI/US/XR)	Level 3 (degraded)	539 cases (300 adjudicated)	Expert-panel adjudication	Yes (report as image)	Bassett, et al. [10]

Note: bpMRI, biparametric MRI; CTPA, CT pulmonary angiography; EHR, electronic health record; US, ultrasound; XR, radiography.

4.1. Single-Volume Datasets (Level 1)

The anchor for chest-CT work is CT-RATE, a paired collection of 25,692 non-contrast 3D chest CT volumes from 21,304 patients with corresponding reports, which underpins the CT-CLIP contrastive model and the CT-CHAT conversational VLM Hamamci, et al. [2]. CT-RATE and its derivatives (for example segmentation-guided VQA and multi-granular retrieval sets built on it) have made chest CT the *de facto* testbed for volumetric vision–language work. Its strengths — scale, paired reports, open access — are real; its granularity is Level 1 by construction, as it provides one series per examination and a report that already encodes the answer.

Generalist 3D resources extend this pattern across regions. RadFM assembled a large multimodal medical dataset spanning 13 million 2D images and 615,000 3D scans and introduced the RadBench evaluation, reporting that its model outperformed then-available multimodal systems including GPT-4V on radiologic tasks [3]. Merlin, an abdominal-CT vision–language foundation model trained on 25,528 CT scans with paired electronic-health-record data and reports, was evaluated across roughly fifty thousand abdominal and chest CTs at internal and external sites and reported large gains over prior radiology foundation models [4]. These are among the most capable volumetric systems in the literature, but their evaluation input remains a single volume per task.

4.2. Multi-Series/Multiparametric Datasets (Level 2)

A minority of resources reach Level 2, providing the several series a radiologist would use for a specific question — but curated to a fixed set for one body part. In liver imaging, MCT-LTDiag provides complete four-phase CT (non-contrast, arterial, portal-venous, delayed) for 517 cases across five tumour categories, explicitly reflecting that hepatic lesion characterisation is a multi-phase task [6]. In prostate imaging, the PI-CAI resource and challenge are built on biparametric MRI and supported an international, paired, non-inferiority comparison of AI and radiologists for clinically significant prostate-cancer detection [7]. In MRI more broadly, fastMRI+ adds subspecialist bounding-box and study-level pathology annotations to fully sampled knee and brain MRI (16,154 box annotations and 13 study-level labels across 22 knee pathologies; 7,570 boxes and multiple study-level labels across 30 brain pathologies), providing the substrate for multi-slice grounded evaluation [8]. These datasets demonstrate that Level-2 evaluation is feasible where the clinical question fixes the relevant series — but each is confined to a single body part and a predetermined sequence/phase set.

4.3. Whole-Study Resources and Their Scarcity (Level 3)

Level-3 resources — the complete examination with all clinically relevant series and priors — are rare. Where whole-study realism appears, it tends to come bundled inside multimodal *clinical* datasets rather than imaging benchmarks. INSPECT links 23,248 CT pulmonary angiography studies from 19,402 patients with report impressions and longitudinal electronic-health-record data, making it one of the few large resources to connect imaging, text, and outcomes for a defined clinical pathway [9]. VALID, a diagnostic-accuracy evaluation of 539 multimodal inpatient cases, supplied radiology (CT, MRI, X-ray, ultrasound) and radiology reports as images alongside laboratory, narrative, and vital-sign data, with expert-panel adjudication of ground truth [10]. These approximate the information a clinician integrates, but at a cost: the imaging is typically degraded (exported images rather than native volumetric series) and — critically — the radiology *report* is present as an input, so the model can read the radiologist's conclusion rather than the scan.

© 2026 by the authors; licensee Online Academic Press, USA

4.4. How Dataset Input Compares with the Radiologist's Input (RQ2)

The per-level gap is stark. At Level 0–1, the model receives a fraction of the pixels and none of the cross-series context a radiologist uses; the search-and-comparison problem is removed, and, where a report is provided, the interpretive problem is removed too. At Level 2, the input matches the radiologist's for a specific question but not for the open-ended examination. Only at Level 3 does the input approximate reality — and Level-3 resources are scarce, degraded, or leak the report. The dominant datasets thus measure something narrower than, and systematically easier than, clinical reading.

5. Benchmarks and Tasks

Table 2 charts representative 2025–2026 benchmarks by task, granularity, models evaluated, metrics, and comparator.

Table 2. Representative CT/MRI benchmarks (2025–2026), charted by task and granularity.

Benchmark	Task	Granularity	Models evaluated	Metrics	Comparator	Ref.
RadLE (Radiology's Last Exam)	Spot diagnosis (multi-modality)	Level 0–1 (web UI)	Frontier models	Accuracy; error taxonomy	Radiologists & trainees	Datta, et al. [11]
DeepTumorVQA	Hierarchical tumour VQA	Level 1 (3D CT)	>30 VLM/agent configurations	Stage-wise accuracy	Across models; tool-augmented	Chen, et al. [12]
CT-SpatialVQA ("Lost in Volume")	Semantic-spatial reasoning	Level 1 (3D CT)	8 3D medical VLMs	Accuracy vs chance	Across models	Monon, et al. [13]
SGMRI-VQA ("Beyond a Single Frame")	Multi-frame grounded VQA	Level 1–2 (MRI, 12–60 slices)	GPT-4o, Gemini 2.5 Pro/Flash + 5 open models	Detection / classification / diagnosis; grounding	Across models	Moukheiber, et al. [14]
3D-RAD	3D CT VQA (6 tasks, incl. longitudinal)	Level 1–2 (multi-temporal)	Medical & general VLMs	Accuracy per task	Across models	Gai, et al. [15]
RADAR	Report discrepancy review	Level 1 (abdominal CT)	Multimodal models	Agreement / severity / edit type	Across models	Sun, et al. [16]
GMAI-MMBench	General medical multimodal VQA	Mixed	50 LVLMs	Accuracy	Across models	Chen, et al. [17]
MIL 3D neuroimage benchmark	Whole-scan classification	Level 1 (3D, whole scan)	MIL / 3D CNN baselines	Classification metrics	Across methods	Harvey, et al. [18]
GPT-5 multimodal medical reasoning	Exam-style multimodal QA	Level 0 (exam images)	Frontier model vs GPT-4o	Reasoning / understanding	Pre-licensed experts	Shansong, et al. [19]

© 2026 by the authors; licensee Online Academic Press, USA

5.1. Classification and Detection

Classification and detection benchmarks are where the slice-versus-volume scoring distinction bites hardest. When abnormality labels are scored at the slice level, and when validation splits are performed slice-wise rather than subject-wise, apparent accuracy can be substantially inflated (Section 7.3). Volume-level and study-level scoring — asking whether the abnormality was found *in the examination*, with subject-level splits — is the more faithful framing, and is increasingly adopted in the stronger 3D benchmarks. A multi-dataset benchmark of multiple-instance learning for 3D neuroimage classification illustrates the methodological stakes: it found that simple mean-pooling over frozen per-slice embeddings matched or outperformed more elaborate attention-based multiple-instance learning and 3D convolutional networks on several tasks, underscoring how much whole-scan classification performance depends on aggregation and evaluation choices rather than on architectural novelty [18].

5.2. Report Generation

Report generation is the most clinically resonant task and the one where metric choice most distorts conclusions. N-gram overlap metrics (BLEU, ROUGE, METEOR) reward surface phrasing and are insensitive to clinically decisive distinctions such as negation or laterality; a report that inverts "no acute intracranial haemorrhage" to "acute intracranial haemorrhage" can score well on n-gram overlap while being clinically catastrophic. Radiology-specific metrics were developed in response: RadGraph-F1 scores agreement over extracted clinical entities and their relations, and is widely treated as the standard for clinical grounding [20], while LLM-as-judge metrics such as GREEN identify and weight clinically significant errors [21], and embedding/entity metrics such as RaTEScore target clinically relevant semantics [22]. The evaluation literature increasingly reports these alongside — or instead of — n-gram scores, but n-gram metrics remain common, and headline improvements framed in BLEU/ROUGE should be read with caution.

5.3. Visual Question Answering and Grounded Reasoning

VQA is the format in which most frontier-versus-specialist comparison now happens, and 2025–2026 produced a series of increasingly demanding 3D benchmarks. 3D-RAD assembles roughly 170,000 entries across six CT VQA task types, including multi-temporal (longitudinal) diagnosis, and reports that existing VLMs — especially medical VLMs — generalise poorly on multi-temporal tasks [15]. DeepTumorVQA decomposes abdominal-tumour reasoning on 9,262 CT volumes (about 476,000 questions) into recognition, measurement, visual reasoning, and medical reasoning, and finds that reliable quantitative *measurement* is the primary bottleneck, with tool augmentation substantially mitigating it [12]. CT-SpatialVQA isolates semantic-spatial reasoning (localisation, laterality, structural comparison, inter-structure relations) over 9,077 QA pairs derived from 1,601 reports and volumes, and reports that eight 3D medical VLMs average only about 34% — frequently below chance — indicating that models often rely on language priors rather than grounded volumetric evidence [13]. In MRI, SGMRI-VQA ("Beyond a Single Frame") builds 41,307 pairs from fastMRI+ that require reasoning across 12–60 sequential slices with frame-indexed bounding boxes, explicitly targeting multi-frame grounded reasoning rather than single-frame classification [14]. The general-medicine benchmark GMAI-MMBench, spanning 284 datasets and 38 modalities, found even a leading frontier model reaching only about 52% accuracy overall, situating radiology within a broader ceiling on medical multimodal reasoning [17].

5.4. Challenge and Leaderboard Efforts

Challenge-based evaluation standardises data splits and hidden test sets in ways single-paper benchmarks cannot. PI-CAI is the clearest recent exemplar: algorithms were ranked on a hidden 1,000-scan cohort, the top entries retrained on a larger set, and the winners compared against radiologists in a paired, non-inferiority design published in a peer-reviewed clinical venue [7]. Newer efforts extend standardisation to report-centric tasks — for example RADAR, which frames report *discrepancy* review (agreement, severity, edit type) on expert-annotated abdominal CT, mirroring the trainee-to-attending revision workflow [16]. What challenges standardise (splits, hidden tests, comparator design) is exactly what many preprint benchmarks leave under-specified; what they still rarely standardise is Level-3 whole-study input.

6. Evaluation of Frontier Versus Specialist Models

6.1. How Frontier Models Are Fed Volumetric Data

Closed frontier models expose an image-and-text API, not a volumetric one. To evaluate them on CT/MRI, investigators must serialise a 3D study into something the API accepts: sampled axial slices, a montage grid, or a "slice-video." The binding constraint is the per-request image budget — for a leading vision API, nominally up to 1,500 images and a 512 MB payload per request, with practical degradation well before that cap and much tighter limits on some enterprise deployments (for example ten images per request) [23]. A single thin-slice CT series can exceed several hundred images, and a whole study several thousand, so faithful whole-study input collides directly with the API envelope. The common workarounds — downsampling to a handful of slices, or providing only a key slice — push frontier-model evaluation toward

© 2026 by the authors; licensee Online Academic Press, USA

Level 0–1 regardless of intent. A frequent further shortcut is to supply the radiology *report* as text and use the frontier model as a reasoner over the report, which measures language reasoning rather than image interpretation and constitutes report leakage (Sections 4.3, 7.3).

6.2. What the Evidence Shows (RQ1, RQ4 in Part)

Three consistent findings emerge from 2025–2026 head-to-head evaluations. First, on genuinely volumetric and measurement-dependent tasks, frontier models underperform. On CT-SpatialVQA, 3D medical VLMs — and the frontier models benchmarked alongside them — cluster near or below chance on semantic-spatial reasoning [13]; on DeepTumorVQA, measurement is the bottleneck that caps later-stage reasoning across more than thirty model configurations [12]. Second, on expert-level diagnostic difficulty, frontier models trail radiologists by a wide margin. In Radiology's Last Exam (RadLE), a pilot of 50 expert-level "spot diagnosis" cases evaluated through native web interfaces, board-certified radiologists scored 83% and trainees 45%, while the best frontier model reached 30% (with other frontier systems at 29%, 23%, 12%, and 1%), and the authors catalogue a taxonomy of visual reasoning errors [11]. Third, the apparent exceptions are informative: where frontier models post strong numbers on medical multimodal *reasoning* benchmarks — for example large gains reported for a newest-generation model on a multimodal medical question-answering benchmark, surpassing pre-licensed human experts [19] — the input is typically exam-style images and text rather than native volumetric studies, and the task rewards structured medical knowledge more than volumetric search. On the imaging tasks that most resemble reading, frontier models are frequently used only as text judges or excluded from the image path entirely.

Specialist 3D models fare better on volumetric benchmarks by design — Merlin, CT-CLIP/CT-CHAT, RadFM, M3D, and efficient successors such as Med3DVLM are trained to ingest volumes — but their advantage is measured largely within Level 1, and even they degrade sharply when evaluation moves to Level-2 multi-series reasoning or grounded spatial tasks [13, 15].

6.3. Failure Modes

The recurring failure modes are coherent across studies: hallucination of findings not present; weak cross-slice and cross-series grounding (answering from language priors rather than the volume) [13]; a measurement bottleneck, where models cannot reliably quantify size, count, or change, which then caps downstream reasoning [12]; and sensitivity to slice sampling, where the same study yields different answers depending on which slices were serialised into the request. The RadLE error taxonomy — spanning perceptual misses, mislocalisation, and flawed reasoning chains — provides a useful shared vocabulary for these behaviours [11].

7. Do the Scores Represent Real Value for Patients and Radiologists?

7.1. Metric-to-Value Mapping (RQ3)

A metric is clinically meaningful to the extent that it tracks a decision. Area under the receiver-operating-characteristic curve (AUROC) summarises discrimination across all thresholds, but clinical value usually lives at a specific operating point — detection *sensitivity at a clinically acceptable specificity*, or the false-positive rate a radiologist will tolerate per study. F1 conflates precision and recall at a single implicit threshold that may not match clinical priorities. For report generation, n-gram metrics (BLEU/ROUGE/METEOR) track phrasing, not correctness; RadGraph-F1, GREEN, and RaTEScore track clinical entities, errors, and semantics and are far better proxies for report quality [20-22]. None of these routinely capture what ultimately matters: whether the output changes management correctly, and whether its errors are safe (a missed incidental hyperdensity and a missed large-vessel occlusion are not equal). The mapping from common metric to clinical decision is therefore partial at best; headline accuracy rarely answers "would this help the radiologist, and would its mistakes hurt the patient?"

7.2. Reference-Standard Quality

Every benchmark is bounded by its reference standard. Report-derived labels inherit the original radiologist's errors and the imprecision of automated label extraction. RECIST key-slice annotations capture a single measured slice, not the lesion in three dimensions or its full burden. Expert panels — as adjudicated in VALID, where independent reviewers resolved discordant cases to establish ground-truth diagnoses, differentials, and reasoning [10] — are stronger but expensive and still subject to reader variability. Outcome- and EHR-anchored labels, as in INSPECT's linkage of imaging to longitudinal records [9] offer the most clinically grounded target but are available for few pathways. A benchmark can never claim more clinical validity than its reference standard supports, and much of the literature is anchored on the weakest of these (report-derived labels).

7.3. The Realism Confounds

Three confounds jointly inflate apparent readiness. *Report leakage:* when the radiology report is available as input, the model can restate the radiologist's conclusion; several whole-study-adjacent resources carry this

© 2026 by the authors; licensee Online Academic Press, USA

flag by design (Sections 4.3, 6.1) [10]. *Curated positive slices:* presenting a key slice on which the finding is visible removes detection and localisation. *Subject-level versus slice-level leakage:* when slices from one patient appear in both training and test sets, accuracy is inflated — quantified in a foundational analysis where slice-wise splitting yielded ~94% apparent accuracy while subject-wise splitting on the same data fell to ~66%, a ~28-percentage-point artefact of the validation protocol alone [24]. Case selection (enriching for clear positives) compounds all three. These confounds are not exotic; they are common, and they push measured performance systematically above what a model would achieve on consecutive, unselected clinical studies.

8. Deployability and Utility: Proprietary Frontier Versus Open/Smaller Models

8.1. The Access Asymmetry (RQ4)

Benchmark performance and deployability point in opposite directions. Closed frontier models are reachable through a maintained, inexpensive API with no local GPU, no model-ops burden, and no need to hold protected health information on bespoke infrastructure beyond the API boundary — but they are weakest on volumetric reading and constrained by image caps (Section 6.1). Open specialist 3D models — RadFM, CT-CLIP/CT-CHAT, Merlin, M3D, Med3DVLM, and open medical VLMs such as MedGemma [25] — post the strongest volumetric numbers but require self-hosting: GPU capacity, security and compliance engineering, versioning, and ongoing maintenance [2-4]. Even parameter-efficient designs intended for clinical deployment (for example Med3DVLM's factorised volumetric encoder [5]) still presuppose an institution willing to run and maintain a model. The practical cost of the high-benchmark option is therefore high, and the practical cost of the low-friction option is a model not yet evaluated for whole-study reading.

8.2. Implication for Evaluation Priorities

This asymmetry has a direct consequence for where evaluation effort should go. The option that is deployable at scale — the API frontier model — is the one least rigorously evaluated on whole studies, precisely because the API envelope pushes it toward Level 0–1. The field is thus over-investing in Level-1 evaluation of models that are hard to deploy, and under-investing in Level-3 evaluation of the models most likely to be deployed. Closing this gap requires realistic, study-level evaluation *of frontier API models specifically*, including the serialisation strategies (slice sampling, montages, agentic tool use) that mediate their access to volumes.

8.3. Cost, Latency, Governance, and Reproducibility

Three further trade-offs shape real-world utility. *Cost and latency:* per-study inference cost and turnaround differ by orders of magnitude between a cloud API call over a handful of slices and a local 3D model over a full volume. *Governance:* sending imaging to an external API raises data-protection questions that a self-hosted model avoids, while a self-hosted model shifts the compliance burden onto the institution. *Reproducibility:* frontier API models are moving targets — versions are deprecated and updated, so a benchmark result attaches to a model *snapshot* and may not replicate months later. Any evaluation intended to inform deployment must record the exact model version, serialisation method, and prompt, and should be read with an expectation of churn (Section 11).

9. The Clinical-Realism Gap (Synthesis, RQ5)

Pulling the threads together, four forces jointly explain why benchmark performance overstates clinical readiness. *Input incompleteness:* evaluation clusters at Levels 0–1, presenting a fraction of the study and removing the search, cross-series integration, and comparison that dominate real reading (Sections 4, 6). *Report leakage:* the resources that approximate whole-study realism frequently supply the radiologist's report as input, converting an interpretation task into a reading-comprehension task (Sections 4.3, 7.3). *Isolated-classification framing:* tasks are posed as "is finding X present?" on curated inputs rather than "read this study and tell me what matters," collapsing the coupled detect–characterise–compare–synthesise operation into a single classification (Section 2.3). *Leakage-inflated metrics:* slice-level scoring and slice-wise splits inflate accuracy by tens of percentage points relative to subject-level, study-level evaluation (Section 7.3) [24].

These are not independent problems that happen to co-occur; they compound. A benchmark that is Level 0–1, supplies the report, frames the task as isolated classification, and scores it slice-wise will report a number far above the model's real diagnostic value — and that number will be the one that shapes adoption decisions. The 2025–2026 benchmarks that deliberately close one or more of these gaps — DeepTumorVQA on measurement [12], CT-SpatialVQA on grounding [13], SGMRI-VQA on multi-frame reasoning [14], 3D-RAD on longitudinal comparison [15], RadLE on expert-level difficulty [11] — are precisely the ones on which frontier and specialist models look weakest. The pattern is consistent: as evaluation moves up the granularity taxonomy and closes realism confounds, measured performance falls toward the clinical truth.

© 2026 by the authors; licensee Online Academic Press, USA

10. Toward Study-Level, Workflow-Faithful Evaluation

10.1. Desiderata

We propose that CT/MRI model evaluation intended to inform clinical adoption should aim for the following, framed as field-level desiderata rather than prescriptions for any product.

1. Whole-study input (Level 3): Provide all clinically relevant series, phases, and planes, plus prior studies where the clinical question requires comparison. Where the model or API cannot ingest a full study, report the serialisation method explicitly and treat it as part of the system under test.
2. Withhold the radiology report: The report is the answer key; it must not be an input to an image-interpretation evaluation. Report availability should be a mandatory disclosed flag.
3. Reference standards tied to consensus or outcomes: Prefer expert-panel adjudication and/or outcome/EHR linkage over single-reader report-derived labels; state the standard and its limits.
4. Task framing that mirrors the read: Evaluate the coupled operation — detect, then localise, then characterise, then compare with priors, then synthesise an impression — not isolated classification on curated slices.
5. Report safety and tail-risk: Report sensitivity at clinically acceptable specificity, the rate and severity of missed critical findings, and hallucination rates, not only aggregate accuracy or AUROC.
6. Deployability-aware evaluation of API models: Because API frontier models are the deployable-at-scale option, evaluate them at study level under their real constraints (image caps, latency, cost, version), and record the exact snapshot.
7. Subject-level, study-level scoring with leakage control: Use subject-wise splits and study-level scoring; audit for train/test contamination and case-selection enrichment.

10.2. A Proposed Evaluation Checklist

Table 3 renders these desiderata as a checklist the community can apply to any CT/MRI model evaluation. It is intended as a shared, non-proprietary instrument: an evaluation that can tick most items is closer to measuring clinical value; one that ticks few is likely measuring something easier than reading.

Table 3. Study-level, workflow-faithful evaluation checklist.

#	Item	What to report	Rationale
1	Input granularity	Level 0/1/2/3; series and planes provided	Higher levels approximate real reading (§2.2)
2	Serialisation method	Slices sampled, montage, slice-video, native volume, or full DICOM; per-request image budget	API caps push frontier models to Level 0–1 (§6.1)
3	Priors included?	Whether prior studies were provided for comparison	Comparison is core to the read (§2.3)
4	Report withheld?	Explicit flag: report available as input (Y/N)	Report leakage converts interpretation into comprehension (§7.3)
5	Reference standard	Report-derived / RECIST key-slice / expert panel / outcome-EHR	Bounds what the benchmark can claim (§7.2)
6	Task framing	Isolated classification vs detect → characterise → compare → synthesise	Determines how much of the read is preserved (§2.3)
7	Data-split integrity	Subject-wise vs slice-wise; contamination audit	Slice-wise splits inflate accuracy (§7.3)
8	Scoring level	Slice-level vs volume/study-level	Study-level scoring is the faithful unit (§5.1)
9	Safety / tail-risk	Sensitivity at clinical specificity; missed-critical rate; hallucination rate	Aggregate accuracy hides unsafe errors (§7.1)
10	Comparator	Radiologist / trainee / other models / none	Human comparator anchors clinical meaning (§6.2)
11	Deployability	API vs self-hosted; cost, latency, governance	Deployable option must be evaluated at Level 3 (§8)
12	Reproducibility	Exact model snapshot, prompt, date	API models are moving targets (§8.3)

© 2026 by the authors; licensee Online Academic Press, USA

10.3. Open Problems

Several hard problems stand between current practice and study-level evaluation. *Feeding full multi-series studies within API limits* remains unsolved; agentic approaches that let a model request specific series or slices, or call measurement and segmentation tools, are promising — tool augmentation relieved the measurement bottleneck in DeepTumorVQA [12] — but add their own evaluation complexity. *Multi-series fusion and grounded cross-series reasoning* — using the interplay of phases or sequences to characterise a lesion — are barely benchmarked above Level 2. *Benchmark governance and leakage control* need standing infrastructure: hidden test sets, subject-level splits, and report-withholding enforced by the benchmark rather than trusted to the submitter, as challenge-based efforts have begun to demonstrate [7, 16]. And *version-stable evaluation of moving-target API models* requires conventions for snapshotting and re-testing that the field has not yet adopted.

11. Limitations

This review has several limitations. The evidence base is heavily weighted toward preprints, which have not completed peer review and whose results and version numbers may change; we verified identifiers and status at write-up, but rapid churn is intrinsic to this field and specific model results may be superseded quickly. The scoping design maps the field rather than exhaustively cataloguing it, and our snapshot of the field should be read accordingly (Section 3.5). We restricted to English-language work and to CT/MRI, excluding modalities (radiography, mammography, ultrasound, pathology) where analogous granularity issues arise. Single-reviewer charting in a rapid review carries a higher risk of omission than dual independent screening; the pre-registered criteria and rejection log mitigate but do not eliminate this. Finally, because one author has a competing interest, we have been deliberately conservative — naming no product or company, deriving recommendations from the charted evidence rather than the reverse, and framing all proposals as field-level desiderata.

12. Conclusion

Current evaluation of large multimodal models on CT and MRI is conducted overwhelmingly at the level of a single slice or a single volume, while a radiology study is a multi-series, multi-plane, multi-sequence object read with priors and clinical context. This granularity mismatch, compounded by report leakage, isolated-classification framing, and leakage-inflated metrics, inflates apparent performance and obscures the failure modes that matter clinically. The models most likely to be deployed at scale — inexpensive, API-accessible frontier systems — are the least rigorously evaluated on whole studies, while the strongest volumetric benchmarks target specialist models that carry a heavy self-hosting burden. Across the 2025–2026 literature, the consistent signal is that as evaluation climbs the granularity taxonomy and closes realism confounds, measured performance falls toward clinical reality. Closing the gap will require study-level, workflow-faithful evaluation — whole-study input, the report withheld, consensus- or outcome-based reference standards, read-mirroring task framing, explicit safety reporting, and deployability-aware testing of API models. We offer the checklist in Table 3 as a step toward that goal. The aim is not to discourage the use of these models, which are advancing quickly, but to measure them against the object radiologists actually read — so that reported progress corresponds to progress patients can feel.

References

[1] Y. Lin, Y. Yang, G. Shih, and Y. Peng, "Rethinking radiology AI benchmarks," *Radiology Advances* vol. 3, no. 3, p. umag022, 2026, https://doi.org/10.1093/radadv/umag022.

[2] I. E. Hamamci *et al.*, "Generalist foundation models from a multimodal dataset for 3D computed tomography," *Nature Biomedical Engineering*, pp. 1-19, 2026, https://doi.org/10.1038/s41551-025-01599-y.

[3] C. Wu, X. Zhang, Y. Zhang, H. Hui, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data," *Nature Communications*, vol. 16, no. 1, p. 7866, 2025, https://doi.org/10.1038/s41467-025-62385-7.

[4] L. Blankemeier *et al.*, "Merlin: A computed tomography vision–language foundation model and dataset," *Nature*, vol. 652, no. 8112, pp. 1318-1328, 2026, https://doi.org/10.1038/s41586-026-10181-8.

[5] Y. Xin, G. C. Ates, K. Gong, and W. Shao, "Med3dvlm: An efficient vision-language model for 3d medical image analysis," *IEEE Journal of Biomedical and Health Informatics*, 2025, https://doi.org/10.48550/arXiv.2503.20047.

[6] X. a. Wu *et al.*, "A multi-phase CT dataset for automated differential diagnosis of liver tumors," *Scientific Data*, vol. 13, no. 1, p. 31, 2025, https://doi.org/10.1038/s41597-025-06343-4.

[7] A. Saha *et al.*, "Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): An international, paired, non-inferiority, confirmatory study," *The Lancet Oncology*, vol. 25, no. 7, pp. 879–887, 2024, https://doi.org/10.1016/S1470-2045(24)00220-1.

[8] R. Zhao *et al.*, "fastMRI+: Clinical pathology annotations for knee and brain fully sampled multi-coil mri data," *Scientific Data*, vol. 9, no. 1, p. 152, 2021, https://doi.org/10.48550/arXiv.2109.03812.

[9] S.-C. Huang *et al.*, "Inspect: A multimodal dataset for pulmonary embolism diagnosis and prognosis," *arXiv preprint arXiv:2311.10798*, 2023, https://doi.org/10.48550/arXiv.2311.10798.

© 2026 by the authors; licensee Online Academic Press, USA

[10] B. A. Bassett *et al.*, "Evaluating multimodal LLMs for inpatient diagnosis: Real-world performance, safety, and cost across ten frontier models," *arXiv preprint arXiv:2604.16980*, 2026, https://doi.org/10.48550/arXiv.2604.16980.

[11] S. Datta *et al.*, "Radiology's last exam (RadLE): Benchmarking frontier multimodal AI against human experts and a taxonomy of visual reasoning errors in radiology," *arXiv preprint arXiv:2509.25559*, 2025, https://doi.org/10.48550/arXiv.2509.25559.

[12] Y. Chen *et al.*, "DeepTumorVQA: A hierarchical 3D CT benchmark for stage-wise evaluation of medical VLMs and tool-augmented agents," *arXiv preprint arXiv:2605.09679*, 2026, https://doi.org/10.48550/arXiv.2605.09679.

[13] M. Monon, U. Rahman, A. Hanif, N. Saeed, and M. Yaqub, "Lost in volume: The CT-spatialVQA benchmark for evaluating semantic-spatial understanding of 3D medical vision-language models," *arXiv preprint arXiv:2605.08787*, 2026, https://doi.org/10.48550/arXiv.2605.08787.

[14] L. Moukheiber, C. M. Yeung, H. Xue, A. Helbling, Z. Zhao, and Y. Chen, "Beyond a single frame: Multi-frame spatially grounded reasoning across volumetric MRI," *arXiv preprint arXiv:2604.15808*, 2026, https://doi.org/10.48550/arXiv.2604.15808.

[15] X. Gai, J. Liu, Y. Li, Z. Meng, J. Wu, and Z. Liu, "3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks," *Advances in Neural Information Processing Systems*, vol. 38, 2026, https://doi.org/10.48550/arXiv.2506.11147.

[16] Z. Sun *et al.*, "RADAR: A multimodal benchmark for 3d image-based radiology report review," *arXiv preprint arXiv:2603.06681*, 2026, https://doi.org/10.48550/arXiv.2603.06681.

[17] P. Chen *et al.*, "Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai," *Advances in Neural Information Processing Systems*, vol. 37, pp. 94327-94427, 2024, https://doi.org/10.48550/arXiv.2408.03361.

[18] E. Harvey, D. J. Loevlie, A. A. Satani, W. Chen, D. M. Kent, and M. C. Hughes, "A multi-dataset benchmark of multiple instance learning for 3D neuroimage classification," *arXiv preprint arXiv:2604.26807*, 2026, https://doi.org/10.48550/arXiv.2604.26807.

[19] W. Shansong, H. Mingzhe, L. Qiang, S. Mojtaba, and Y. Xiaofeng, "Capabilities of GPT-5 on multimodal medical reasoning," *arXiv:2508.08224*, N.D.

[20] F. Yu *et al.*, "Evaluating progress in automatic chest x-ray radiology report generation," *Patterns*, vol. 4, no. 9, p. 100802, 2023, https://doi.org/10.1016/j.patter.2023.100802.

[21] S. Ostmeier *et al.*, "GREEN: Generative radiology report evaluation and error notation," presented at the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 2024, pp. 374–390, https://doi.org/10.18653/v1/2024.findings-emnlp.21, 2024.

[22] W. Zhao, C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "RaTEScore: A metric for radiology report generation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[23] OpenAI, "OpenAI API documentation: Image and file inputs," OpenAI, Retrieved: https://platform.openai.com/docs. 2024.

[24] E. Yagis *et al.*, "Effect of data leakage in brain MRI classification using 2D convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, p. 22544, 2021, https://doi.org/10.1038/s41598-021-01681-w.

[25] A. Sellergren *et al.*, "Medgemma technical report," *arXiv preprint arXiv:2507.05201*, 2025, https://doi.org/10.48550/arXiv.2507.05201.

© 2026 by the authors; licensee Online Academic Press, USA