Review

Trustworthiness Evaluation of Medical Vision-Language Models: Scoping Review and Proposed Minimum Evaluation Checklist

Binesh Sadanandan[1], MS, PhD; Amirhossein Karimi[1], MS; Bibek Upadhayay[2], PhD; Vahid Behzadan[1], PhD

[1]University of New Haven, SAIL Lab, West Haven, CT, United States
[2]Saginaw Valley State University, University Center, MI, United States

Corresponding Author:

Binesh Sadanandan, MS, PhD
University of New Haven, SAIL Lab
300 Boston Post Rd
West Haven, CT 06516
United States
Phone: 1 (203) 932-7000
Email: bsada1@unh.newhaven.edu

Abstract

Background: Medical vision-language models combine images and text to answer questions, generate reports, and support image interpretation. Task accuracy alone does not establish whether their outputs depend on relevant visual evidence or remain reliable under changes in the input.

Objective: This study aimed to describe how trustworthiness is evaluated in medical vision-language models, identify gaps in evaluation and reporting, and develop a preliminary minimum evaluation checklist.

Methods: We conducted a scoping review using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews). We searched PubMed/MEDLINE, Embase, Scopus, Web of Science Core Collection, and IEEE Xplore for English-language publications from January 1, 2022, to March 25, 2026. Eligible reports were peer-reviewed full papers that evaluated an image-and-language model in a medical imaging context and reported a measured result on at least 1 of 8 trustworthiness dimensions. One reviewer charted the studies, and a second verified the extraction. We summarized evaluation practices descriptively and applied the proposed 8-item MiTEC (Minimum Trustworthiness Evaluation Checklist) retrospectively. We did not perform a risk-of-bias appraisal.

Results: We identified 516 records, screened 506 database records based on the title and abstract, and assessed 72 unique reports after merging duplicate full-text entries. The revised synthesis includes 29 studies: 11 journal articles and 18 conference papers. Robustness was evaluated in 37.9% (11/29), hallucination in 24.1% (7/29), distribution shift in 20.7% (6/29), and visual grounding in 17.2% (5/29). Interpretability and fairness were each evaluated in 13.8% (4/29), uncertainty or confidence reliability in 6.9% (2/29), and safety in 3.4% (1/29). Two studies used a text-only baseline, and 4 used at least 1 image-reliance control. No included study was confirmed to report a formal calibration metric or a selective-prediction curve. Retrospective MiTEC scoring identified 41 (17.7%) reported items among 232 study-item pairs.

Conclusions: Evaluation practices varied across the included studies. Image-reliance controls, formal calibration measures, and subgroup analyses were uncommon. MiTEC is an author-proposed item set that requires independent testing and stakeholder consensus. The findings describe the included sample; incomplete linkage of preprints to published versions limits claims about the wider literature.

JMIR AI 2026;5:e102330; doi: 10.2196/102330

Keywords: AI; deep learning; diagnostic imaging; vision-language model; multimodal large language model; trustworthiness; reproducibility of results; uncertainty; calibration; scoping review

Introduction

Background

Vision-language models (VLMs) combine image representations with natural language. In medicine, they are used for visual question answering (VQA), report generation, image classification, and localization of findings. Medical adaptations include LLaVA-Med, Med-Flamingo, and BiomedCLIP, while general-purpose systems have also been evaluated on clinical images [1-4]. These systems differ in architecture and training data, but they raise a shared evaluation problem: a correct answer does not, by itself, show that a model has used the image appropriately.

Medical VQA benchmarks commonly compare answers with a reference label or short text. Report-generation studies often use lexical overlap or automated measures of clinical content. Such measures are useful for comparing task performance, but they address only part of the evidence needed to judge a model's behavior. A system may answer from question priors, produce unsupported findings, or change its answer when the wording of a clinically equivalent question changes. Confidence can also be misleading when it does not correspond to observed correctness [5-8].

Image dependence deserves particular attention in a multimodal model. When a question can be answered without the image, high accuracy may reflect associations in the text. SURE-VQA (Systematic Understanding of Robustness Evaluation in Medical VQA) tested baselines that ignored visual input and showed why they matter when interpreting medical VQA results [9]. Yu et al [10] compared radiology report generators with image-independent retrieval baselines, providing a related control in report generation. These peer-reviewed studies support examining image reliance separately from answer accuracy.

The clinical meaning of other trustworthiness properties also depends on the evaluation design. Robustness concerns the response to an input change; calibration concerns agreement between confidence and correctness. Fairness requires examination of differences between relevant patient groups. An explanation may be plausible without faithfully representing the process that produced the answer [11]. Treating these properties as interchangeable can obscure weaknesses in the evidence.

Previous Reviews and Review Objectives

Previous reviews have described medical VLM architectures, imaging applications, and the development of generative AI in medicine [12-15]. Other work has examined trustworthiness in health care language models or the quality of medical training data [16,17]. These contributions provide useful context, but their organizing questions differ from a study-level account of evaluation procedures.

Benchmarks have addressed several trustworthiness dimensions within the same experiment. CARES (Comprehensively Evaluate the Trustworthiness of Medical Large Vision Language Models), for example, evaluates medical VLMs across multiple properties, including fairness, safety, and robustness [18]. We therefore do not claim to introduce multidimensional trustworthiness evaluation. Our contribution is a literature synthesis of which properties are evaluated, how those evaluations are conducted, and which controls and reporting practices accompany them.

We addressed 6 questions: what studies and clinical tasks have been evaluated; which trustworthiness dimensions have been measured; what evaluation methods and metrics have been used; which controls test image reliance; what mitigation strategies or deployment safeguards have been reported; and what reporting gaps remain. We used these findings to develop a preliminary MiTEC (Minimum Trustworthiness Evaluation Checklist). A scoping review was appropriate because the objective was to map methods across heterogeneous tasks rather than estimate a pooled treatment or diagnostic effect.

Methods

Review Design and Eligibility

We followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; Checklist 1) [19]. The protocol was not prospectively registered. We have not deposited a retrospective protocol and do not claim a registration identifier.

We included peer-reviewed journal articles and full conference or workshop papers published from January 1, 2022, to March 25, 2026. Reports had to be published in English, concern medical or clinical imaging, evaluate a model involving images and natural language, and report a measured result on at least 1 of the 8 dimensions in Table 1. Task accuracy alone was insufficient. A mechanism intended to improve trustworthiness, a proposed evaluation design, or illustrative figures without a measured outcome did not satisfy the criterion. We applied this rule without exceptions for publication type or anticipated relevance.

Table 1. Operational scope and rationale for the 8 trustworthiness dimensions.

Dimension	Operational definition	Rationale
Robustness	Measured change under a controlled image or text perturbation.	Named by CARES[a] and Aljohani et al [16]; separates controlled input changes from clinical transfer.
Hallucination	Measured unsupported or factually incorrect output, including factuality in generated reports.	Maps to trustfulness in CARES and truthfulness in Aljohani et al.

Dimension	Operational definition	Rationale
Distribution shift	Performance or detection under a stated change in data distribution.	Separates generalization across settings from a fixed-set perturbation.
Visual grounding	Measured connection between language and the image or a relevant image region.	Adds the image-dependence question specific to multimodal evaluation.
Interpretability	Reported evaluation of explanations, concepts, or attribution.	Maps to explainability in Aljohani et al; faithfulness requires separate evidence.
Fairness	Measured differences between demographic groups or a bias-reduction evaluation.	Named by CARES and Aljohani et al.
Uncertainty or confidence reliability	Measured relationship between an uncertainty signal and correctness. Formal calibration is coded separately.	Adds the medical imaging distinction between error detection and probability calibration [20,21].
Safety	Measured harmful or unsafe output, including resistance to elicitation of such output.	Named by CARES and Aljohani et al; diagnostic error alone is coded elsewhere.

[a]CARES: Comprehensively Evaluate the Trustworthiness of Medical Large Vision Language Models.

We excluded text-only language models, image-only systems, nonmedical applications, reviews, commentaries, protocols, conference abstracts, and preprints without a peer-reviewed full-paper version assessed for the synthesis. Background citations and the supplementary literature scan are separate from the included sample. We also excluded author-affiliated primary studies to limit conflicts arising from assessment of our own work. Multimedia Appendix 1 records the available exclusion reasons and the 5 eligibility corrections made during revision.

Trustworthiness Dimensions and Coding Rules

The 8 dimensions were chosen to describe the evaluation of model outputs and their dependence on clinical inputs. They were informed by the medical AI data-quality framework of Schwabe et al [17], the health care language-model survey of Aljohani et al [16], and the properties evaluated by CARES [17,18]. The mapping in Table 1 states the scope of each dimension. We separated visual grounding from factual correctness because an answer can be correct without using the relevant image. We also separated distribution shift from controlled input perturbation, while allowing a study to address more than one dimension when distinct results justified the codes.

Privacy was outside the prespecified scope, and the search terms did not target privacy threats. Evaluating membership disclosure, memorization, or access to sensitive records requires additional threat models and data-access assumptions. CARES includes privacy as a separate property; its inclusion in that benchmark does not make the 8 dimensions here an exhaustive account of trustworthiness.

We distinguished uncertainty evaluation from formal calibration. The former includes an assessment of whether an uncertainty signal identifies errors. The latter requires a measure such as expected calibration error, Brier score, negative log-likelihood used to assess calibration, or a reliability diagram. Discrimination measures and entropy values alone do not establish calibration.

We also distinguished text perturbation from paraphrase consistency. A text perturbation changes an input and measures its effect on the output. A paraphrase-consistency test holds clinical meaning fixed, presents alternative wordings of the same question, and reports item-level agreement or answer changes. Replacing a disease term changes meaning. Comparing a reasoning question with an entailed subquestion also differs from rephrasing the same question. These distinctions were applied when reviewing the charting and are defined in Multimedia Appendix 2.

Information Sources and Search Strategy

Binesh Sadanandan searched PubMed/MEDLINE, Embase through Ovid, Scopus, Web of Science Core Collection, and IEEE Xplore on March 15, 2026, with an update on March 25, 2026. The searches used the strings in the originally submitted Multimedia Appendix 3 and were restricted to English. Results were imported into Zotero for reference management and deduplication, then into Rayyan for screening [22]. Backward and forward citation searching of included reports and 6 prior reviews, together with targeted update searches, contributed additional records for full-text assessment.

The strategies combined model terms, medical imaging terms, and trustworthiness terms. Alternatives within each concept were joined with "OR," and the 3 concepts were joined with "AND." No "NOT" block was used. Multimedia Appendix 3 gives every database-specific string, field tag, date limit, language restriction, and reported publication-type limit. It reproduces the original strategies separately from the syntax corrections made during revision and explains phrase matching and truncation for each platform.

During revision, we checked the corrected strategies on the platforms available to us. PubMed searches on July 29, 2026, returned 848 records with the submitted string and 870 with the revised string within the stated publication window. Three revised IEEE Xplore searches on August 17, 2026, returned 496 unique records; the annual filter also retrieved records published after the cutoff. We did not screen these later retrievals into this review. They are reported separately in Multimedia Appendix 3 and do not replace the March screening set.

Study Selection

Binesh Sadanandan and Amirhossein Karimi screened titles and abstracts. Rayyan contains decisions from both reviewers for 461 records and 1 recorded decision for 45. Disagreements were resolved by discussion, with Vahid Behzadan acting as the adjudicating reviewer where needed. Author-affiliated studies were excluded under the same policy regardless of which reviewer identified them. This reduces direct self-assessment, although the authors' related research may still have influenced the questions emphasized in the review.

We report the Cohen κ for the 461 paired decisions. The previously reported value of 0.91 treats a "maybe" decision as "not include" and compares "include" with "not include." Treating "include," "exclude," and "maybe" as 3 categories gives a κ of 0.47. We state the coding convention because the 2 estimates answer different questions. Under the Landis and Koch convention, the binary estimate indicates almost perfect agreement, and the 3-category estimate indicates moderate agreement [23]. These labels should be interpreted with the decision coding in mind.

Full reports were assessed against the eligibility criteria. During revision, we reexamined the studies whose publication type or dimension coding was in question. Shah and OmniMedVQA were checked against the published full texts, including supplementary material. We removed these 2 studies, the Khatri conference abstract, the Xian commentary, and EviVLM. The original identifiers remain visible in the correction record; they were not reassigned to other studies.

Data Charting and Synthesis

We piloted a charting form on 10 studies. Binesh Sadanandan completed the extraction, and Amirhossein Karimi verified the charted records against the available publications. Disagreements were resolved by discussion. This was extraction with verification rather than 2 independent extractions. We did not retain independent charting decisions from which to calculate a charting-stage agreement statistic.

Multimedia Appendix 2 contains the study table, definitions of all 36 fields, and the record of eligibility and coding changes. The fields cover bibliographic information, clinical context, model type, the 8 dimensions, 5 image-reliance controls, and reporting indicators. Additional fields distinguish uncertainty evaluation from formal calibration and broad text perturbation from paraphrase consistency. Datasets, mitigation methods, and some task-specific metrics were summarized narratively because they were not separate charting fields.

He et al [24] reliability study was a published conference paper, but we could not obtain its full text for this review. Its robustness and text-variation coding rests on the reported evaluation described in the abstract. The paraphrase-consistency field is undetermined. Other negative entries for that study mean that the available abstract did not report the practice; they should not be interpreted as a full-text confirmation of absence.

We calculated counts and percentages from the revised 29-study table. Each study could contribute to several dimensions, and the count of any image-reliance control is the union of the 5 control fields. We did not perform a quality or risk-of-bias appraisal, weight studies by size, or pool performance estimates across tasks. The reported proportions describe the frequency of evaluation practices.

Development and Retrospective Application of MiTEC

We developed MiTEC from the gaps identified in the review and compared its items with existing AI reporting guidance. The checklist was proposed by the review team; it was not developed through a Delphi process or external consensus exercise. We retrospectively scored all included studies against explicit pass rules, using the charted fields and additional source checks for image perturbations, paraphrase consistency, and selective prediction.

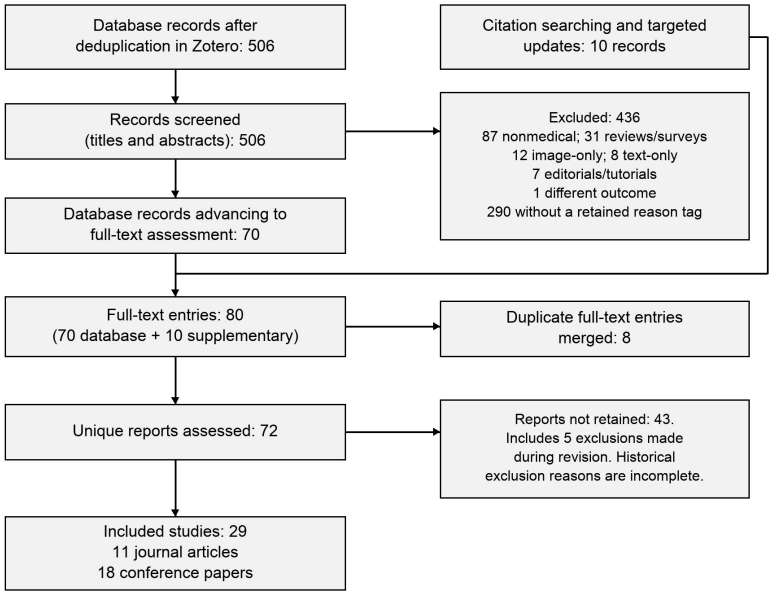
The retrospective analysis assesses whether the specified practice was reported. It does not measure overall study quality. Unclear evidence received no credit in the primary score and remained identified separately. Multimedia Appendix 4 gives every item rule, study-level score, supporting evidence, and sensitivity analysis. For the reproducibility item, the primary score uses reported code availability because a data-availability statement was not consistently charted. A sensitivity analysis requires code and data. No claim of checklist validation follows from this exercise.

Results

Selection of Studies

We identified 516 records: 506 database records and 10 from citation searching and targeted updates. Of the 506 database records screened at the title and abstract levels, 436 were excluded and 70 advanced. The 10 supplementary records entered at full-text assessment. Merging 8 duplicate full-text entries left 72 unique reports, of which 29 were included and 43 were not retained in the revised synthesis (Figure 1).

Figure 1. Study selection. The revised synthesis contains 29 studies after 5 eligibility corrections. Total identified: 516. Multimedia Appendix 5 contains a separate 28-record literature list; it is not part of the full-text flow.



Exclusion-reason tags were available for 146 title-and-abstract exclusions: 87 concerned nonmedical work, 31 were reviews or surveys, 12 lacked a language component, 8 were text-only models, 7 were editorials or tutorials, and 1 concerned a different outcome. The remaining 290 had no retained reason tag. These categories account for all 436 exclusions; the tagged subset does not establish the distribution of reasons among untagged records.

The original manuscript reported 34 included studies. Applying the stated eligibility rule during revision removed 5: Shah, Xian, Khatri, OmniMedVQA, and EviVLM. Multimedia Appendix 1 gives the reason and source for each decision. SURE-VQA and Med-VCD (visual-contrastive decoding) remain included and have rows in Multimedia Appendix 2. Full-text reason records are incomplete, so we do not infer a verified preprint-exclusion subtotal from the difference between the flow counts. Multimedia Appendix 5 contains 28 distinct supplementary literature entries with publication-status notes. It is a curated literature list, not a list of all reports excluded at full text.

Study Characteristics

The sample comprised 11 journal articles and 18 conference papers. Publication years are shown in Table 2. We retained the bibliographic years recorded in the charting table, which use issue years for some journal articles and online publication years for others. Multimedia Appendix 2 identifies these cases. SURE-VQA uses its 2025 Transactions on Machine Learning Research publication year. The 2026 count covers publications up to March 25.

Table 2. Publication years of the included peer-reviewed studies (N=29).

Publication year	Studies, n (%)
2022	3 (10.3)
2023	2 (6.9)
2024	6 (20.7)
2025	17 (58.6)
2026	1 (3.4)

Seventeen studies evaluated general and medical-specific models, 11 evaluated medical-specific models, and 1 evaluated a general-purpose model. Nineteen used the broad multiple-specialty category. Seven were restricted to radiology, 1 to pathology, and 2 combined named specialties. Six studies were coded as chest radiograph only, 1 as histology, 2 as named modality combinations, and 20 as multiple modalities.

VQA was the most common named task, followed by classification. Report generation, localization, and out-of-distribution detection were also represented. Broad labels such as multiple modalities and multiple model families limit finer comparisons; they should not be interpreted as showing that all models or modalities received equivalent testing. No included study evaluated trustworthiness during an open-ended, multiturn medical VLM consultation. This finding concerns our sample rather than the absence of relevant work in the wider literature.

Trustworthiness Dimensions Evaluated

Frequency of Evaluated Dimensions

The most frequently evaluated dimension was robustness, followed by hallucination and distribution shift (Table 3). Most studies evaluated 1 dimension (21/29, 72.4%). The references in Table 3 identify every study contributing to each count; Multimedia Appendix 2 provides the individual codes.

Table 3. Trustworthiness dimensions evaluated among the 29 included studies.[a]

Dimension	Studies, n (%)	All contributing studies
Robustness	11 (37.9)	[7,9,18,24-31]
Hallucination	7 (24.1)	[7,10,18,27,32-34]
Distribution shift	6 (20.7)	[9,26,31,35-37]
Visual grounding	5 (17.2)	[29,38-41]
Interpretability	4 (13.8)	[42-45]
Fairness	4 (13.8)	[18,46-48]
Uncertainty or confidence reliability	2 (6.9)	[18,32]
Safety	1 (3.4)	[18]

[a]All contributing studies are cited; categories overlap.

Robustness and Distribution Shift

Eleven studies evaluated robustness. Cheng et al [26] tested image artifacts across medical imaging tasks, while Mozhegova et al [25] and Han et al [30] examined adversarial changes to images or paired text. HEAL-MedVQA measured answer changes after replacement of disease or anatomy terms and intervention on the queried image region [29]. These are measured perturbation results, although the changed questions are not equivalent paraphrases.

Seven studies applied a text perturbation and reported its effect. The changes included synonym substitution or deletion, instruction format, answer-option wording, added context, adversarial captions, entity substitution, and question variation. The text-variation result in He et al [24] was charted from its abstract. No study was confirmed to report item-level consistency across clinically equivalent paraphrases; that property remained undetermined for their study. Tascon-Morales et al [28] measured consistency between logically related questions, which addresses a different relation.

Six studies were charted for distribution shift. The designs included artifact-related changes, cross-task or cross-domain evaluation, out-of-distribution detection, and adaptation across medical VQA datasets [9,26,31,35-37]. Lai et al [35] and Ju et al [36] treated shift detection as an explicit task. SURE-VQA evaluated predefined shift axes and used sanity baselines to help interpret the results. These designs address different aspects of generalization and should not be read as equivalent tests of clinical transportability.

Hallucination and Visual Grounding

Seven studies evaluated hallucination or factual correctness. MedVH assessed several forms of unsupported medical output, while Huppertz et al [27] examined fabricated findings in GPT-4V responses [7]. Liao et al [32] assessed an uncertainty signal for hallucination detection. Med-VCD evaluated a decoding method intended to reduce hallucination [34]. Delbrouck et al [33] and Yu et al [10] examined factuality in radiology report generation; CARES included factual errors within its wider evaluation [18].

Five studies evaluated visual grounding. Their approaches included localization before answering, parameter-efficient visual grounding, and outputs tied to image regions or pixels [29,38,39,41]. Sun et al [40] examined unimodal bias in medical VQA. Grounding scores and localization outputs provide information about the relationship between language and image regions. They do not automatically establish that an answer changes appropriately when the relevant image evidence changes.

We use grounding here to mean visual grounding. Grounding an answer in retrieved documents or cited evidence is a separate property. The distinction matters when comparing image-based tasks with conversational clinical systems.

Uncertainty, Fairness, Interpretability, and Safety

Two studies evaluated uncertainty or confidence reliability. Liao et al [32] reported discrimination measures for a hallucination-detection signal, and CARES reported uncertainty-based accuracy and an overconfidence ratio [18]. No included study was confirmed to report a formal calibration metric. A useful ranking of uncertain answers does not establish that stated confidence corresponds to the probability of a correct answer.

Four studies evaluated demographic fairness and reported subgroup analyses [18,46-48]. The studies examined performance differences between groups or methods intended to reduce them. Yang et al [46] also considered intersectional groups. The recorded subgroup analyses were limited relative to the range of populations in which medical VLMs may be used.

Four studies were charted for interpretability, using attention, concepts, or natural-language explanations [42-45]. Their reported explanation methods do not provide

interchangeable evidence of faithfulness. An attention map or a readable rationale can help describe a model's output while leaving its causal basis uncertain. We identified no included study applying sparse autoencoder or circuit-level analysis to a medical image-and-language evaluation.

CARES was the only study charted for safety as a distinct measured outcome [18]. We reserved this code for harmful or unsafe outputs and resistance to prompts intended to elicit them. Adversarial image or text changes scored only by diagnostic error were counted under robustness. This distinction prevents the same term, adversarial, from determining the code regardless of what was measured.

Evaluation Methods and Image-Reliance Controls

VQA-RAD, SLAKE, PathVQA, and chest radiograph datasets were used across the included work. Evaluation methods varied with the task. Classification and VQA studies reported answer-level measures; report-generation studies used lexical or clinical-content measures; grounding studies used localization measures. These metrics address different end points, and their numerical values cannot be pooled into one trustworthiness score.

Only 4 studies reported at least 1 of the 5 image-reliance controls (Table 4). Yu et al [10] used image-independent retrieval baselines for report generation. SURE-VQA used baselines that ignored image input [9]. These 2 studies provide the in-sample evidence for the value of a text-only comparison.

Table 4. Controls for image reliance.[a]

Control	Studies, n (%)	Contributing studies
Text-only or image-independent baseline	2 (6.9)	[9,10]
Image swap	1 (3.4)	[7]
Region-of-interest intervention	1 (3.4)	[29]
Counterfactual image or text	1 (3.4)	[29]
Saliency checked against relevant regions	0 (0)	None
Any control	4 (13.8)	Four unique studies
No control	25 (86.2)	Complement of the union

[a]A study can contribute to more than 1 control; the final rows count unique studies.

MedVH replaced the question image with an unrelated image and assessed whether models recognized the mismatch [7]. HEAL-MedVQA replaced the queried image region with a corresponding region from a different class and changed disease or anatomy terms [29]. Its regional intervention and counterfactual tests therefore appear in separate rows, while the study is counted once in the union. Neither the presence of a localization output nor a saliency image alone was treated as a control for image dependence.

Mitigation Methods and Reporting Practices

Reported mitigation methods included adversarial fine-tuning, factuality-oriented training, visual contrastive decoding, grounding before answer generation, and fairness-oriented learning [29,30,33,34,47,48]. The studies evaluated these methods within different tasks and datasets. We did not compare their effect sizes across studies.

Nine studies reported external evaluation under the charting definition, which included evaluation on data separate from training or development. This is broader than prospective validation at an independent clinical site. Eighteen studies reported code availability; 5 were coded no and 6 remained unknown. Six reported a data release. Table 5 gives the reporting indicators with the denominator stated for each.

Table 5. Evaluation and reporting practices in the included sample (N=29).[a]

Practice	Studies, n (%)	Interpretation
External evaluation	9 (31.0)	Separate from training or development; not necessarily an independent clinical site
Demographic subgroup analysis	4 (13.8)	All 4 also address fairness
Code availability reported	18 (62.1)	5 no; 6 unknown
Data released	6 (20.7)	Study-specific data or annotations
Formal calibration metric	0 (0)	Distinct from uncertainty discrimination
Train-test leakage check	0 (0)	Explicit contamination assessment

Practice	Studies, n (%)	Interpretation
Text perturbation evaluated	7 (24.1)	Includes meaning-changing perturbations
Equivalent-question consistency confirmed	0 (0)	1 undetermined study, S20
Selective-prediction curve confirmed	0 (0)	Within this included sample only

[a]Undetermined evidence is identified separately from a confirmed positive code.

No included study was confirmed to report a selective-prediction curve. This finding does not establish that abstention methods were absent from the field. In particular, a precutoff peer-reviewed version of the Wienholt study was not linked to the screened record during the review. Its omission is discussed below.

Discussion

Principal Findings

The included studies evaluated a wider range of properties than task accuracy, but the coverage was uneven. Robustness and hallucination received the most attention. Formal calibration metrics, image-independent baselines, and subgroup analyses were less common. Four studies used at least 1 image-reliance control, leaving 25 without a reported control under our definitions.

Image dependence and answer reliability need to be considered together. A model may answer consistently because it follows a text prior. Conversely, a model may respond to an image intervention while remaining poorly calibrated. The reviewed procedures often addressed one part of this problem. SURE-VQA and the image-independent report-generation baselines show how an apparently simple control can change the interpretation of performance [9,10].

The distinction between uncertainty and calibration was also consequential. An uncertainty method can rank incorrect answers above correct answers without assigning well-calibrated probabilities. The 2 included uncertainty studies provide evidence about their respective signals, but they do not fill the formal-calibration gap. Evaluation should state what the confidence quantity represents and how its relationship to correctness was measured.

Our proportions describe the frequency of reported practices. They do not show how thoroughly those practices were implemented, and they do not estimate the clinical reliability of the evaluated models. A study with a small experiment contributes equally to a dimension count as one with a large benchmark. The lack of a quality appraisal makes that limitation particularly relevant when interpreting headline percentages.

Comparison With Previous Work

Tables 6 and 7 compare this review with the prior reviews, frameworks, and benchmarks discussed in the Introduction. The comparison distinguishes literature synthesis from evaluation of models within a benchmark. Coverage grades summarize the scope of each work; they are not the binary study codes used in Multimedia Appendix 2.

Table 6. Scope and dimensional coverage of related work.[a]

Work	Study type and scope	Coverage of the 8 dimensions
Hartsock and Rasool [12]	Review: medical VLMs[b]; 16 models, 18 datasets	Partial: interpretability.
Buess et al [13]	Scoping review: generative AI in medicine; 145 studies	Partial: distribution shift, visual grounding, interpretability, fairness, uncertainty/calibration. Not reported: safety.
Ryu et al [14]	Review: medical imaging VLMs; 61 papers	Partial: distribution shift, interpretability, fairness.
Schouten et al [15]	Scoping review: multimodal AI; 432 papers	Partial: distribution shift, interpretability.
Aljohani et al [16]	Survey: health care text-only language models	Full: robustness, hallucination, interpretability, fairness, safety.
Schwabe et al [17]	Review-derived data-quality framework; 120 records	None of the 8 model-output dimensions.
Xia et al [18]	Benchmark: medical VLMs; about 41,000 question-answer pairs	Full: robustness, hallucination, fairness, safety. Partial: distribution shift, uncertainty/calibration.
Gu et al [7]	Benchmark: medical hallucination; about 2700 items	Full: hallucination. Partial: robustness, visual grounding.
Chang et al [49]	Preprint benchmark: hallucination; 11 models	Full: hallucination.
This review	Scoping review: evaluation practices; 29 studies	Full: robustness, hallucination, distribution shift, visual grounding, interpretability, fairness, uncertainty/calibration, safety.

[a]"Full" indicates systematic treatment within the work's own scope; "partial" indicates narrative or limited treatment. Privacy lies outside this review's 8 dimensions.
[b]VLM: vision-language model.

Table 7. Treatment of image-reliance controls, uncertainty, fairness, and reporting gaps in related work.[a]

Work	Image controls	Uncertainty/calibration	Fairness	Reporting gaps
Hartsock and Rasool [12]	No image-reliance control synthesis	Not addressed	Not addressed	No
Buess et al [13]	No image-reliance control synthesis	Single-study discussion	Dataset-bias discussion	No
Ryu et al [14]	No image-reliance control synthesis	Not addressed	Dataset-bias discussion	No
Schouten et al [15]	No image-reliance control synthesis	Not addressed	Not addressed	External-validation and missing-data reporting
Aljohani et al [16]	Not applicable to image input	Not a survey dimension	Survey dimension	No cross-study reporting-gap analysis
Schwabe et al [17]	Data-quality framework	Not addressed	Data-side bias	Data-quality awareness framework
Xia et al [18]	Image corruptions; no image-omission or image-swap condition	Uncertainty accuracy and overconfidence; no formal calibration metric	Age, sex, and race	Benchmark, not a literature synthesis
Gu et al [7]	Wrongful-image task	False-confidence task tests rejection of a suggested wrong answer	Not addressed	Benchmark, not a literature synthesis
Chang et al [49]	Not reported	Not reported	Not reported	Preprint benchmark
This review	4/29 (13.8%) with any charted control	2/29 (6.9%) uncertainty; 0 formal calibration	4/29 (13.8%) demographic fairness	Study-level reporting indicators and proposed MiTEC[b]

[a]MedHEval is a preprint and is included here for context only.
[b]MiTEC: Minimum Trustworthiness Evaluation Checklist.

Dimensions not listed for a work were not addressed in this comparison. CARES and Aljohani et al [16] additionally address privacy. Schwabe et al [17] address data quality rather than model-output dimensions. Full coverage describes the organizing scope; it does not imply that every property received a distinct measured test.

CARES already demonstrated the value of evaluating several trustworthiness properties together. MedVH concentrated on medical hallucination and included a wrongful-image task. The contribution of this review is to compare evaluation practices across studies and make the definitions underlying the counts explicit.

Related clinical language-model work also helps delimit the findings. Hu [50] proposed retrieval-grounded evaluation of conversational risk assessment, including citation faithfulness and subgroup audits. That proposal addresses evidence grounding in a text-based system, whereas our grounding codes concern images. Hu [51] provides a discussion of clinical reasoning that further illustrates why an explanation can be persuasive while resting on inappropriate evidence. These reports inform interpretation but are not included studies under the image-and-language eligibility rule.

The MIRAGE preprint raises a related concern about answers generated without visual input [52]. We treat it as supplementary, unreviewed evidence. The peer-reviewed image-independent controls in the sample provide the principal basis for recommending such comparisons.

Proposed Minimum Evaluation Checklist

MiTEC comprises 8 proposed items (Table 8). Several adapt established reporting expectations to medical VLMs. Calibration, subgroup analysis, external evaluation, and access to research materials have clear precedents in clinical AI guidance [53-57]. Text-only baselines, image interventions, and equivalent-question tests specify procedures that are particularly relevant to image-and-language systems. We do not claim that these procedures are new methods.

Table 8. Proposed MiTEC (Minimum Trustworthiness Evaluation Checklist) items and their relationship to existing reporting expectations.[a]

Item	Proposed evaluation or reporting practice	Relationship to existing guidance
Text-only baseline	Report performance with image input withheld or an image-independent comparison.	VLM[b]-specific procedure; broader guidance describes inputs but does not establish image dependence.
Image intervention	Change, mask, corrupt, or replace an image and measure the effect on the answer.	VLM-specific test related to the input-quality concerns in CONSORT-AI[c], SPIRIT-AI[d], and DECIDE-AI[e].
Equivalent questions	Test at least 3 clinically equivalent phrasings of each question and report agreement or answer changes.	VLM-specific application of consistency testing; the proposed threshold requires testing.
Calibration	Report a formal calibration metric or reliability diagram for a stated confidence quantity.	Adaptation of calibration reporting in TRIPOD+AI[f]. An uncertainty score alone is insufficient.
Subgroups	Report performance by relevant demographic groups and the observed gaps.	Adaptation of fairness and subgroup reporting in TRIPOD+AI, STARD-AI[g], and DECIDE-AI.

Item	Proposed evaluation or reporting practice	Relationship to existing guidance
External evaluation	Evaluate a separate dataset and describe its relation to development data and any performance change.	Adaptation of external-evaluation expectations in STARD-AI and TRIPOD+AI.
Selective prediction	Report a risk-coverage curve or its area for confidence-based abstention.	Specific uncertainty safeguard; related guidance addresses indeterminate or poor-quality inputs.
Research materials	Describe access to code, data, and model weights; provide a data-availability statement.	Adaptation of access and reproducibility reporting in TRIPOD+AI and CONSORT-AI.

[a]These are author-proposed evaluation items, not a consensus standard.
[b]VLM: vision-language model.
[c]CONSORT-AI: Consolidated Standards of Reporting Trials-AI.
[d]SPIRIT-AI: Standard Protocol Items: Recommendations for Interventional Trials-AI.
[e]DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence.
[f]TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis+AI.
[g]STARD-AI: Standards for Reporting Diagnostic Accuracy—AI.

The comparison concerns the stated procedures, rather than a claim of method novelty. The retrospective scoring uses narrower observable proxies for items 6 and 8, as specified in Multimedia Appendix 4.

We propose a threshold of at least 3 equivalent phrasings as a practical proposal for an initial consistency test. The review does not establish that 3 is sufficient. The number and diversity of variants should be examined during prospective testing of the checklist. Similarly, the choice of subgroup and external dataset should reflect the intended clinical setting.

Selective prediction is a plausible safeguard because it allows a model to withhold answers when uncertainty is high. Lu et al [58] demonstrated uncertainty-aware selective classification in clinical-trial prediction, an adjacent clinical AI task. That example supports feasibility in principle; it is not evidence that the same procedure has been validated for medical VLM use. A risk-coverage analysis should report the trade-off between error among retained predictions and the proportion of cases receiving an answer.

Retrospective Checklist Findings

The retrospective analysis identified 41 (17.7%) reported items among 232 study-item pairs. The median score was 1 of 8, with an interquartile range of 1 to 2 and a range of 0 to 4. Seven studies received no credit under the primary rules. Table 9 gives the item-level results.

Table 9. Retrospective application of MiTEC (Minimum Trustworthiness Evaluation Checklist) to the 29 included studies.[a]

Item	Studies meeting the retrospective rule, n (%)
Text-only baseline	2 (6.9)
Image perturbation or swap test	8 (27.6)
Prompt paraphrase consistency	0 (0)
Calibration metric	0 (0)
Subgroup stratification	4 (13.8)
External or out-of-distribution evaluation	9 (31.0)
Selective prediction curve	0 (0)
Reproducibility artifacts	18 (62.1)

[a]One paraphrase-consistency cell remained undetermined and received no credit in the primary score.

The scores depend on the pass rules. Counting unknown code-release status as a pass increased the total to 47 of 232 (20.3%) items. Requiring code and data for the reproducibility item reduced it to 29 (12.5%). Restricting the image-intervention item to a literal image swap reduced it to 34 (14.7%). Crediting the undetermined paraphrase cell increased it to 42 (18.1%). Multimedia Appendix 4 gives the full matrix and evidence for each decision.

This application makes the checklist's interpretation inspectable, but the same team developed and applied it to the same sample. The exercise does not establish content validity, inter-rater reliability, or usefulness in clinical deployment. Independent scoring, prospective pilots, and a consensus process involving clinicians, patients, methodologists, and model developers are needed before the item set could be proposed as a reporting standard.

Limitations

The review has limitations in study identification. We did not consistently link preprints to peer-reviewed versions available before the March 25, 2026, cutoff. Wienholt and colleagues' European Radiology article, published online on February 20, 2026, is one example. VividMed and MedVQA-APL also have precutoff proceedings records among material previously classified as preprints. Nakka's workshop paper appeared before the cutoff, but its screening route is unresolved and the evaluated pipeline uses image features without a language input. We have not added selected records during this revision: resolving publication versions

consistently would require rechecking the full candidate set. These omissions can affect the prevalence estimates, particularly the statements about selective prediction and grounding. Multimedia Appendix 5 identifies the relevant publication-status corrections.

The later PubMed and IEEE Xplore query checks were not fully screened and could not be reconciled with the original combined database import. Their yields cannot be used as revised screening counts. We retain the original reported screening totals and state that the historical full-text reason record is incomplete. Multimedia Appendix 1 provides the reasons available to us, rather than presenting a reconstructed reason for every excluded report. English-language eligibility, inconsistent terminology, and the fixed search cutoff may also have limited retrieval.

Charting involved 1 extractor and a verifying reviewer. We did not retain independent charting decisions, and no agreement statistic can be calculated for that stage. The reliability study by He et al [24] was charted from its abstract, which limits the detail and certainty of its codes. Several other fields use broad categories that do not support detailed comparisons of model versions, patient populations, or datasets. We did not undertake a formal quality or risk-of-bias appraisal, so the dimension counts treat studies as equally weighted observations regardless of sample size or evaluation depth.

The selected dimensions omit privacy and other aspects of clinical implementation. Author-affiliated studies were excluded, but our research interests may still have influenced the review's scope. Finally, MiTEC is an unvalidated proposal. Its retrospective scores should be interpreted as a description of reporting under stated rules, not as certification of a model or a study.

Conclusions

Among the 29 included studies, robustness and hallucination were evaluated more often than formal calibration, subgroup performance, or image dependence. The findings support clearer definitions and explicit controls when interpreting medical VLM results. MiTEC offers a preliminary set of evaluation items for further testing. Independent assessment and stakeholder consensus are needed before adoption as a reporting standard, and the review's study-identification limitations should be considered when applying its findings.

Acknowledgments

The authors thank the Secure and Assured Intelligent Learning Lab at the University of New Haven for computing infrastructure and literature-database access.

Generative AI Use

Claude (Opus 4.6; Anthropic) was used for language editing during preparation of the earlier manuscript. Claude and OpenAI Codex were used during revision for language editing, checking references and numerical consistency, source retrieval, and preparation of the revision files. The authors are responsible for study selection, charting decisions, interpretation, and the final manuscript.

Funding

The authors declared no financial support was received for this work. Institutional computing and library resources were provided as routine departmental support.

Data Availability

Multimedia Appendices 1-5 and Checklist 1 accompany the revised manuscript. They contain the search strategies and their corrections, the charting table and codebook, eligibility and coding changes, the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist, available exclusion reasons, the supplementary literature list, and the retrospective MiTEC (Minimum Trustworthiness Evaluation Checklist) rules and study-level results. The Rayyan export and later search exports are available from the corresponding author on request. No public data deposit has been made.

Authors' Contributions

BS contributed to conceptualization, methodology, screening, data curation, analysis, software, and original drafting. AK contributed to screening, charting verification, validation, and revision. BU advised on methodology and checklist development and contributed to interpretation and revision. VB contributed to conceptualization, methodology, adjudication, resources, supervision, and revision.

Conflicts of Interest

BS is employed as a research engineer at Medtronic Surgical, Connecticut, United States. His role has no association with medical imaging. Medtronic had no role in the design, conduct, funding, or reporting of the review. BS and VB have conducted related primary research on medical vision-language model evaluation; author-affiliated studies were excluded from this synthesis. AK declares no competing interests. BU declares no competing interests.

Multimedia Appendix 1

Excluded studies.
[DOCX File (Microsoft Word File), 25 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Charting data.
[DOC File (Microsoft Word File), 173 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Search strategies.
[DOCX File (Microsoft Word File), 30 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Retrospective MiTEC rules, study-level evidence, scores, and sensitivity analyses.
[PNG File (Portable Network Graphics File), 518 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Horizon scan.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 5]

Checklist 1

PRISMA-ScR checklist.
[PDF File (Adobe File), 36 KB-Checklist 1]

References

1. Li C, Wong C, Zhang S, et al. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc (NeurIPS); 2023:28541-28564. [doi: 10.52202/075280-1240]
2. Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, et al. Med-flamingo: a multimodal medical few-shot learner. In: Hegselmann S, Parziale A, Shanmugam D, Tang S, Asiedu MN, Chang S, Hartvigsen T, Singh H, editors. Proceedings of the 3rd Machine Learning for Health Symposium. Proceedings of Machine Learning Research (PMLR); 2023:353-367. URL: https://proceedings.mlr.press/v225/moor23a.html [Accessed 2026-11-04]
3. Zhang S, Xu Y, Usuyama N, et al. A multimodal biomedical foundation model trained from fifteen million image-text pairs. NEJM AI. Jan 2025;2(1):AIoa2400640. [doi: 10.1056/AIoa2400640] [Medline: 42932457]
4. Li Y, Liu Y, Wang Z, et al. A comprehensive study of GPT-4V's multimodal capabilities in medical imaging. medRxiv. Preprint posted online on Nov 4, 2023. [doi: 10.1101/2023.11.03.23298067]
5. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA matter: elevating the role of image understanding in visual question answering. Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017; Honolulu, HI. [doi: 10.1109/CVPR.2017.670]
6. Agrawal A, Batra D, Parikh D, Kembhavi A. Don't just assume; look and answer: overcoming priors for visual question answering. Presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 18-23, 2018; Salt Lake City, UT. [doi: 10.1109/CVPR.2018.00522]
7. Gu Z, Chen J, Liu F, Yin C, Zhang P. MedVH: toward systematic evaluation of hallucination for large vision language models in the medical context. Adv Intell Syst. Jan 2026;8(1):2500255. [doi: 10.1002/aisy.202500255]
8. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research (PMLR); 2017:1321-1330. URL: https://proceedings.mlr.press/v70/guo17a.html [Accessed 2026-11-04]
9. Kahl KC, Erkan S, Traub J, Lüth CT, Maier-Hein K, Maier-Hein L, et al. SURE-VQA: systematic understanding of robustness evaluation in medical VQA tasks. Trans Mach Learn Res. 2025. URL: https://openreview.net/forum?id=qjNdGpgpV8 [Accessed 2026-11-04]
10. Yu F, Endo M, Krishnan R, et al. Evaluating progress in automatic chest X-ray radiology report generation. Patterns (N Y). 2023;4(9). [doi: 10.1016/j.patter.2023.100802] [Medline: 37720336]
11. Jacovi A, Goldberg Y. Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL); 2020:4198-4205. [doi: 10.18653/v1/2020.acl-main.386]

12. Hartsock I, Rasool G. Vision-language models for medical report generation and visual question answering: a review. Front Artif Intell. 2024;7:1430984. [doi: 10.3389/frai.2024.1430984] [Medline: 39628839]
13. Buess L, Keicher M, Navab N, Maier A, Tayebi Arasteh S. From large language models to multimodal AI: a scoping review on the potential of generative AI in medicine. Biomed Eng Lett. 2025;15(5):845-863. [doi: 10.1007/s13534-025-00497-1] [Medline: 40917142]
14. Ryu JS, Kang H, Chu Y, Yang S. Vision-language foundation models for medical imaging: a review of current practices and innovations. Biomed Eng Lett. 2025;15(5):809-830. [doi: 10.1007/s13534-025-00484-6] [Medline: 40917147]
15. Schouten D, Nicoletti G, Dille B, et al. Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications. Med Image Anal. Oct 2025;105:103621. [doi: 10.1016/j.media.2025.103621] [Medline: 40482561]
16. Aljohani M, Hou J, Kommu S, Wang X. A comprehensive survey on the trustworthiness of large language models in healthcare. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V, editors. Findings of the Association for Computational Linguistics: EMNLP 2025. Association for Computational Linguistics (ACL); 2025:6720-6748. [doi: 10.18653/v1/2025.findings-emnlp.356]
17. Schwabe D, Becker K, Seyferth M, Klaß A, Schaeffter T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. NPJ Digit Med. Aug 3, 2024;7(1):203. [doi: 10.1038/s41746-024-01196-4] [Medline: 39097662]
18. Xia P, Chen Z, Tian J, et al. CARES: a comprehensive benchmark of trustworthiness in medical vision language models. In: The Thirty-Eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track. Neural Information Processing Systems Foundation, Inc; 2024:140334-140365. [doi: 10.52202/079017-4455]
19. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. Oct 2, 2018;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
20. Bernhardt M, De Sousa Ribeiro F, Glocker B. Failure detection in medical image classification: a reality check and benchmarking testbed. Trans Mach Learn Res. 2022. URL: https://openreview.net/forum?id=VBHuLfnOMf [Accessed 2026-11-04]
21. Zhang R, Gatsonis C, Steingrimsson JA. Role of calibration in uncertainty-based referral for deep learning. Stat Methods Med Res. May 2023;32(5):927-943. [doi: 10.1177/09622802231158811] [Medline: 37011026]
22. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. Dec 5, 2016;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159-174. [Medline: 843571]
24. He S, Ren D, Pan H, Zhang K, Li Q. Advancing reliable medical VQA: evaluation and enhancement of model robustness. Presented at: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 15-18, 2025; Wuhan, China. [doi: 10.1109/BIBM66473.2025.11356296]
25. Mozhegova E, Khattak AM, Khan A, Garaev R, Rasheed B, Anwar MS. Assessing the adversarial robustness of multimodal medical AI systems: insights into vulnerabilities and modality interactions. Front Med (Lausanne). 2025;12:1606238. [doi: 10.3389/fmed.2025.1606238] [Medline: 40776910]
26. Cheng Z, Ong AY, Wagner SK, et al. Understanding the robustness of vision-language models to medical image artefacts. NPJ Digit Med. Nov 27, 2025;8(1):727. [doi: 10.1038/s41746-025-02108-w] [Medline: 41309994]
27. Huppertz MS, Siepmann R, Topp D, et al. Revolution or risk?-Assessing the potential and challenges of GPT-4V in radiologic image interpretation. Eur Radiol. Mar 2025;35(3):1111-1121. [doi: 10.1007/s00330-024-11115-6] [Medline: 39422726]
28. Tascon-Morales S, Márquez-Neila P, Sznitman R. Consistency-preserving visual question answering in medical imaging. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII. Springer; 2022:386-395. [doi: 10.1007/978-3-031-16452-1_37]
29. Nguyen D, Ho MK, Ta H, et al. Localizing before answering: a benchmark for grounded medical visual question answering. In: Kwok J, editor. Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, {IJCAI-25}. International Joint Conferences on Artificial Intelligence Organization; 2025:7670-7678. [doi: 10.24963/ijcai.2025/853]
30. Han X, Jin L, Ma X, Liu X. Light-weight fine-tuning method for defending adversarial noise in pre-trained medical vision-language models. In: Chen YN, Bansal M, Chen YN, editors. Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics (ACL); 2024:10784-10799. [doi: 10.18653/v1/2024.findings-emnlp.633]
31. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. NEJM AI. Mar 2024;1(3):AIoa2300138. [doi: 10.1056/AIoa2300138] [Medline: 42932741]

32. Liao Z, Hu S, Zou K, Fu H, Zhen L, Xia Y. Vision-amplified semantic entropy for hallucination detection in medical visual question answering. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part V. Springer; 2025:669-679. [doi: 10.1007/978-3-032-04971-1_63]
33. Delbrouck JB, Chambon P, Bluethgen C, Tsai E, Almusa O, Langlotz C. Improving the factual correctness of radiology report generation with semantic rewards. In: Goldberg Y, Kozareva Z, Zhang Y, editors. Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics (ACL); 2022:4348-4360. [doi: 10.18653/v1/2022.findings-emnlp.319]
34. Mahdavi Z, Khodakaramimaghsoud Z, Khaloo H, et al. Med-VCD: mitigating hallucination for medical large vision language models through visual contrastive decoding. Comput Biol Med. Jan 1, 2026;200:111347. [doi: 10.1016/j.compbiomed.2025.111347] [Medline: 41319470]
35. Lai R, Lu X, Chen K, Chen Q, Zheng WS, Wang R. Hierarchical vision-language learning for medical out-of-distribution detection. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part V. Springer; 2025:230-239. [doi: 10.1007/978-3-032-04971-1_22]
36. Ju L, Zhou S, Zhou Y, et al. Delving into out-of-distribution detection with medical vision-language models. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part V. Springer; 2025:133-143. [doi: 10.1007/978-3-032-04971-1_13]
37. Jia Q, Liu S, Chen M, Li T, Yang J. ECSA: mitigating catastrophic forgetting and few-shot generalization in medical visual question answering. Tomography. Oct 20, 2025;11(10):115. [doi: 10.3390/tomography11100115] [Medline: 41150172]
38. He J, Li P, Liu G, Zhong S. Parameter-efficient fine-tuning medical multimodal large language models for medical visual grounding. Presented at: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI); Apr 14-17, 2025; Houston, TX, USA. [doi: 10.1109/ISBI60581.2025.10981029]
39. Chen Y, Xu D, Huang Y, et al. MIMO: a medical vision language model with visual referring multimodal input and pixel grounding multimodal output. Presented at: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 10-17, 2025; Nashville, TN, USA. [doi: 10.1109/CVPR52734.2025.02303]
40. Sun Z, Harit A, Cristea AI, Yu J, Moubayed NA, Shi L. Is unimodal bias always bad for visual question answering? A medical domain study with dynamic attention. Presented at: 2022 IEEE International Conference on Big Data (Big Data); Dec 17-20, 2022; Osaka, Japan. [doi: 10.1109/BigData55660.2022.10020791]
41. Huang X, Huang H, Shen L, et al. A refer-and-ground multimodal large language model for biomedicine. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part XII. Springer; 2024:399-409. [doi: 10.1007/978-3-031-72390-2_38]
42. Naseem U, Khushi M, Kim J. Vision-language transformer for interpretable pathology visual question answering. IEEE J Biomed Health Inform. Apr 2023;27(4):1681-1690. [doi: 10.1109/JBHI.2022.3163751] [Medline: 35358054]
43. Hossain MI, Zamzmi G, Mouton P, Sun Y, Goldgof D. Enhancing concept-based explanation with vision-language models. Presented at: 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS); Jun 26-28, 2024; Guadalajara, Mexico. [doi: 10.1109/CBMS61543.2024.00044]
44. Wang P, Wu J. Large vision-language models-based human-understandable explanations for medical image analysis. Presented at: 2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML); Nov 21-23, 2025; Chongqing, China. [doi: 10.1109/ICICML67980.2025.11333439]
45. Kamal MS, Nimmy SF, Alharbi W. Explainable vision-language model for personalized medicine. In: WWW '25: Companion Proceedings of the ACM on Web Conference 2025. Association for Computing Machinery; 2025:1908-1915. [doi: 10.1145/3701716.3717738]
46. Yang Y, Liu Y, Liu X, et al. Demographic bias of expert-level vision-language foundation models in medical imaging. Sci Adv. Mar 28, 2025;11(13):eadq0305. [doi: 10.1126/sciadv.adq0305] [Medline: 40138420]
47. Wang P, Tong L, Wu J, Liu J, Liu Z. Fair-MoE: medical fairness-oriented mixture of experts in vision-language models. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2025: 28th International Conference, Daejeon, South Korea, September 23–27, 2025, Proceedings, Part V. Springer; 2025:186-196. [doi: 10.1007/978-3-032-04971-1_18]
48. Xue J, Wang B, Wang P, Niu D, Wei B. Mitigating bias catastrophic inheritance in medical large vision-language models with logit fairness adjustment. Presented at: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 15-18, 2025; Wuhan, China. [doi: 10.1109/BIBM66473.2025.11356416]
49. Chang A, Huang L, Bhatia P, Kass-Hout T, Ma F, Xiao C. MedHEval: benchmarking hallucinations and mitigation strategies in medical large vision-language models. arXiv. Preprint posted online on Mar 4, 2025. [doi: 10.48550/arxiv.2503.02157]

50. Hu Y. Toward retrieval-grounded evaluation for conversational large language model-based risk assessment. JMIR AI. Mar 12, 2026;5:e90759. [doi: 10.2196/90759] [Medline: 41818631]
51. Hu Y. Comment on "Classifying the clinical significance of common breast pain symptoms using a large language model, ChatGPT (GPT-4)". Clin Imaging. Apr 2026;132:110741. [doi: 10.1016/j.clinimag.2026.110741] [Medline: 41671902]
52. Asadi M, O'Sullivan JW, Cao F, et al. MIRAGE: the illusion of visual understanding. arXiv. Preprint posted online on Mar 23, 2026. [doi: 10.48550/arXiv.2603.21687]
53. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. May 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
54. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. Oct 2025;31(10):3283-3289. [doi: 10.1038/s41591-025-03953-8] [Medline: 40954311]
55. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
56. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. Sep 2020;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
57. Cruz Rivera S, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. Sep 2020;26(9):1351-1363. [doi: 10.1038/s41591-020-1037-7] [Medline: 32908284]
58. Lu Y, Chen T, Hao N, Van Rechem C, Chen J, Fu T. Uncertainty quantification and interpretability for clinical trial approval prediction. Health Data Sci. 2024;4:0126. [doi: 10.34133/hds.0126] [Medline: 38645573]

Abbreviations

CARES: Comprehensively Evaluate the Trustworthiness of Medical Large Vision Language Models
Med-VCD: visual-contrastive decoding
MiTEC: Minimum Trustworthiness Evaluation Checklist
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SURE-VQA: Systematic Understanding of Robustness Evaluation in Medical Visual Question Answering
VLM: vision-language model
VQA: visual question answering

Edited by Ivan Steenstra; peer-reviewed by Sudip Phuyal, Yihan Hu; submitted 25.May.2026; initial comments to author 29.Jul.2026; final revised version received 02.Oct.2026; accepted 26.Oct.2026; published 25.Nov.2026

Please cite as:
Sadanandan B, Karimi A, Upadhayay B, Behzadan V
Trustworthiness Evaluation of Medical Vision-Language Models: Scoping Review and Proposed Minimum Evaluation Checklist
JMIR AI 2026;5:e102330
URL: https://ai.jmir.org/2026/1/e102330
doi: 10.2196/102330

© Binesh Sadanandan, Amirhossein Karimi, Bibek Upadhayay, Vahid Behzadan. Originally published in JMIR AI (https://ai.jmir.org), 25.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on https://www.ai.jmir.org/, as well as this copyright and license information must be included.