Diagn Interv Radiol 2026; DOI: 10.4274/dir.2026.263812

ARTIFICIAL INTELLIGENCE AND INFORMATICS

GUIDELINE

Reporting checklist for foundation and large language models in medical research (REFINE): an international consensus guideline

Ismail Mese[1], Tugba Akinci D'Antonoli[2,3], Christian Bluethgen[4,5], Keno Bressem[6,7], Renato Cuocolo[8], Akshay Chaudhari[5,9], Ali S. Tejani[10], Amanda Isaac[11], Andrea Ponsiglione[12], Aymen Meddeb[13], Bardia Khosravi[14,15], Bastien Le Guellec[16], Charles E. Kahn, Jr.[17], Chong Hyun Suh[18], Daniel Pinto dos Santos[19], Dow-Mu Koh[20,21], Eleftherios Tzanis[22], Elmar Kotter[23], Errol Colak[24,25], Felipe Kitamura[26], Felix Busch[6], Felix Nensa[27], Guang Yang[28], Henning Müller[29-31], Jakob Nikolas Kather[32-35], Jawed Nawabi[13], Jens Kleesiek[27], Jingyu Zhong[36,37], João Santinha[38,39], Johannes Haubold[40], José Guilherme de Almeida[41], Karim Lekadir[42,43], Kostas Marias[44,45], Lara Noelle Reiner[13], Lena Maier-Hein[46-50], Linda Moy[51], Lisa C. Adams[6], Luis Martí-Bonmatí[52], Magdalini Paschali[5], Mana Moassefi[15], Matthias Dietzel[53], Merel Huisman[54], Michael Ingrisch[55-57], Michail E. Klontzas[22,45], Nikolaos Papanikolaou[58,59], Oliver Diaz[42], Paulo Kuriki[60], Philipp Seeböck[61,62], Pouria Rouzrokh[14,15], Quirin D. Strotzer[63], Seong Ho Park[18], Shahriar Faghani[64,65], Soroosh Tayebi Arasteh[5,66,67], Su Hwan Kim[6], Vasantha Kumar Venugopal[68], Woojin Kim[69-71], Burak Kocak[72]

PURPOSE

To develop the **RE**porting checklist for **F**oundat**I**on and large la**N**guag**E** models (REFINE), an international reporting guideline for transparent and reproducible reporting of foundation model (FM) and large language model (LLM) studies in medical research, including imaging artificial intelligence (AI) applications.

METHODS

The protocol was prespecified and publicly archived. A modified Delphi process was conducted to establish reporting standards for unimodal and multimodal FM and LLM applications involving text, imaging, and structured data. The steering committee coordinated protocol development, expert recruitment, all Delphi rounds, and the harmonization phase. Decisions were made based on predefined consensus thresholds. In Rounds 1 and 2, structured ratings and free-text feedback informed iterative revisions. In the post-Delphi harmonization phase, terminology was standardized, and detailed reporting instructions were finalized.

RESULTS

The REFINE development group comprised 57 contributors from 17 countries, and 54 panelists from 16 countries completed Rounds 1 and 2. The harmonization phase was completed by three expert panelists and the steering committee. The entire process produced a 44-item, six-section framework with standardized terminology and detailed reporting instructions, supported by an online platform for practical use (https://refinechecklist.github.io/refine/checklist.html).

CONCLUSION

The REFINE provides a comprehensive, consensus-based reporting standard for medical FM and LLM research, including imaging AI studies. The online version facilitates practical implementation.

CLINICAL SIGNIFICANCE

The REFINE enables transparent, comparable, and reproducible reporting of FM and LLM studies, supporting reliable evidence synthesis in medical and imaging-focused AI studies.

KEYWORDS

Foundation models, large language models, artificial intelligence, reporting guidelines, medical imaging, Delphi consensus

Corresponding author: Burak Kocak

E-mail: drburakkocak@gmail.com

The authors' affiliations are listed at the end of the article.

Received 18 December 2025; revision requested 11 January 2026; accepted 01 February 2026.

Epub: 26.02.2026

Publication date:

DOI: 10.4274/dir.2026.263812

You may cite this article as: Mese I, Akinci D'Antonoli T, Bluethgen C, et al. Reporting checklist for foundation and large language models in medical research (REFINE): an international consensus guideline. *Diagn Interv Radiol*. 26 February 2026 DOI: 10.4274/dir.2026.263812 [Epub Ahead of Print].

The rapid integration of foundation models (FMs) and large language models (LLMs) into medicine, ranging from complex diagnostics to patient triage,[1,2] is outpacing the scientific community's capacity to conduct rigorous evaluation. These concerns are amplified by the opaque and stochastic behavior of these systems, which limits the applicability of traditional reporting guidelines and contributes to the growing challenge of ensuring reproducibility.

Although several meta-analyses have evaluated LLMs in healthcare, their reliability is limited by fragmented and inconsistent reporting.[3-7] The lack of standardized methodologies and reporting practices, combined with the proprietary black-box nature of these systems, makes comparison of findings challenging.[3,7,8]

FMs and LLMs require distinct reporting standards because their behavior depends on factors that are largely not captured in traditional checklists. These include sensitivity to prompting strategies,[9-11] training dataset specification (e.g., knowledge cutoffs),[12] and the stochastic nature of output generation (e.g., influenced by temperature).[13,14] Furthermore, the scale of these models requires stronger governance regarding intended use, safety, and bias.[15]

To address these gaps, this paper introduces the **RE**porting checklist for **F**oundat**I**on and large la**N**guag**E** models (REFINE) in medical research (Figure 1). The REFINE is a consensus-based checklist that provides clear, item-level guidance to support rigorous reporting and critical appraisal of FM- and LLM-based generative artificial intelligence (AI) studies in medical research, including imaging-focused studies.

Main points

- The **RE**porting checklist for **F**oundat**I**on and large la**N**guag**E** models (REFINE) is an international Delphi-based reporting guideline for studies that use foundation models (FMs) and large language models (LLMs) in medical research.
- The guideline covers six domains: model specification, prompt design, stochasticity control, dataset integrity, output evaluation, and implementation.
- The REFINE items capture critical risks and dependencies inherent to FMs and LLMs that are not entirely addressed in previous reporting frameworks.
- The REFINE is supported by an open, easy-to-use, and multifunctional online platform (https://refinechecklist.github.io/refine/checklist.html).
- Using the REFINE can improve the transparency, reproducibility, and critical appraisal of FM and LLM studies for all key stakeholders, including authors, reviewers, and journal editors.

Methods

Study design

The REFINE was developed using a modified Delphi process. A steering committee (IM, TAD, and BK) developed the protocol and initial set of items, coordinated panel recruitment, and conducted all Delphi rounds and the harmonization phase.

The prespecified protocol, including voting rules, consensus thresholds, and round closure criteria, was deposited on the Open Science Framework before recruitment and was followed without significant deviation. It can be accessed via the following reference.[16]

Scope definition

The steering group defined the scope to develop reporting standards for FMs and LLMs in medical research. Both unimodal and multimodal applications, including text-only, imaging, and structured data studies, are within the scope. The principal intended users of the REFINE are researchers who design, conduct, report, and assess studies involving these models, including authors, reviewers, and editors across medical fields.

Initial item development

First, a review of the relevant literature, including guidelines and methodological works, was conducted.[17-28] Based on this review, an initial item set was drafted, refined for clarity, and organized into distinct sections. This initial item set was used for Round 1.

Panel selection and recruitment

Experts were selected to ensure broad representation across clinical imaging, machine learning, FM and LLM development, medical informatics, methodology, and editorial domains. Invitations were sent directly via email and briefly outlined the aims of the REFINE, the Delphi process, and the co-authorship criteria. Email addresses were used strictly for recruitment and were not linked to survey response data to ensure anonymity.

Anonymity and consent

Each panelist received a unique code to maintain anonymity during voting. These codes enabled tracking of participation while keeping individual responses anonymous. Consent was implied through the entry of

Figure 1. Key features of the REFINE reporting guideline. The QR code links to the online REFINE platform. (https://refinechecklist.github.io/refine/checklist.html).

the code and the submission of responses. No email addresses were collected. Responses were stored securely and used exclusively for the REFINE project.

Consensus criteria and decision rules

Panelists rated each item as "keep as is," "keep with modification," "remove," or "unsure." "Unsure" responses did not count toward consensus. Consensus to keep an item required at least 75% of panelists selecting either "keep as is" or "keep with modification." If one-third or more of these votes indicated "keep with modification," the item was revised according to panelists' comments. Consensus to remove an item required at least 75% of panelists selecting "remove." Items without consensus, as well as those meeting the keep threshold but exceeding the modification threshold, were revised and re-rated in the next round. Items still lacking consensus after Round 2 were removed.

New items were added if proposed by at least two panelists or by one panelist with steering group approval.

Free-text comments were collected for each item, each section, and at the end of Rounds 1 and 2 to inform potential item revisions.

In all other procedural decisions, the steering committee acted by majority vote.

Modified Delphi procedure

Stage 1 (preparation)

The steering group refined the initial items and section structure and tested the survey internally before distributing it to the panelists.

Stage 2 (voting rounds and harmonization phase)

Round 1 (the first formal Delphi round): All items were presented to the entire panel via Google Forms. Panelists provided ratings and free-text comments. The round remained open for 2 weeks, with extensions permitted to maintain adequate participation.

Round 2 (the second formal Delphi round): Items that did not reach consensus, items that reached consensus but required revision based on Round 1 feedback, and any newly proposed items were re-rated. In this round, panelists were also asked to indicate which response options the final checklist should include: i) Yes, No, and N/A or ii) Yes, Partial, No, and N/A. The round remained open for another 2 weeks, with extensions permitted to maintain adequate participation.

Post-Delphi harmonization phase: Following Round 2, the steering committee drafted reporting instructions for each item and invited a small expert group (CB, KB, and RC) from the panel to review them and provide revisions when needed. Under the direction of the steering committee, this group resolved remaining issues, finalized item placement and wording, and established standardized terminology through discussion. This stage produced the final checklist. This phase took place in Google Docs and remained open for 2 weeks.

Statistical analysis

Responses were summarized using descriptive statistics, including proportions meeting the prespecified consensus thresholds. No additional complex statistical analyses were required.

Results

Expert panel characteristics and participation

A total of 55 experts were invited, of whom 54 participated in the Delphi voting rounds, representing 16 countries and multiple disciplines. Including the three steering committee members, the REFINE development group comprised 57 contributors from 17 countries. The combined group composition reflects a high concentration of expertise in radiology-driven AI (68%) and participants predominantly from Germany and the United States (51%), as detailed in Figures 2 and 3.

In Round 1, 54 panelists submitted complete ratings. In Round 2, the same 54 panelists participated. No withdrawals occurred while the rounds were open.

Item evolution

The initial draft included 39 items across five sections. In Round 1, all items met the consensus threshold. Three exceeded the modification threshold and required re-voting; one of these was split into two, yielding four items for re-evaluation. Panel feedback also led to editorial refinements and several new item proposals.

Round 2 evaluated 13 items in total: the four re-evaluation items and nine new proposals. A new section was added, and items were reassigned accordingly. All 13 items achieved consensus, followed by further editorial adjustments and expanded instructional text.

Across the rounds, some consensus items were split into distinct items or combined into a single item to improve clarity.

The harmonization phase finalized the checklist structure, item names and wording, and detailed reporting instructions while maintaining the six-section framework established in Round 2.

Terminology and definitions established and used in the REFINE

To reduce ambiguity in the reporting of FMs and LLMs, the steering committee and the selected expert group established a set of standardized terms during the harmonization phase. These terms describe key stages of model development and evaluation. The standardized terminology is presented in Table 1.

Final REFINE structure

The final REFINE checklist contains 44 items across six sections (model specification, prompt design, stochasticity control, dataset integrity, output evaluation, and implementation). Table 2 provides the complete REFINE checklist. Figure 4 summarizes the consensus statistics for all finalized REFINE items.

Each item includes concise but detailed reporting instructions to support consistent reporting. These instructions clarify intent and provide practical guidance for authors. Table 3 presents the full set of item-level reporting instructions.

The response set used in the final checklist (Yes, Partial, No, and N/A) reflects the preference expressed by the absolute majority of panelists during Round 2.

Web version of the REFINE

A mobile-compatible online version of the REFINE is available at https://refinechecklist.github.io/refine/checklist.html. This version is practical to use and is the recommended format. It integrates the content presented in Tables 2 and 3 by linking each item to its reporting instructions through a tooltip. The online version also provides a real-time summary of completion by section and overall completion. Users can print the checklist to PDF for submission along with their manuscript, export the data as an Excel table for use in systematic reviews, and download the summary statistics image for presentation of

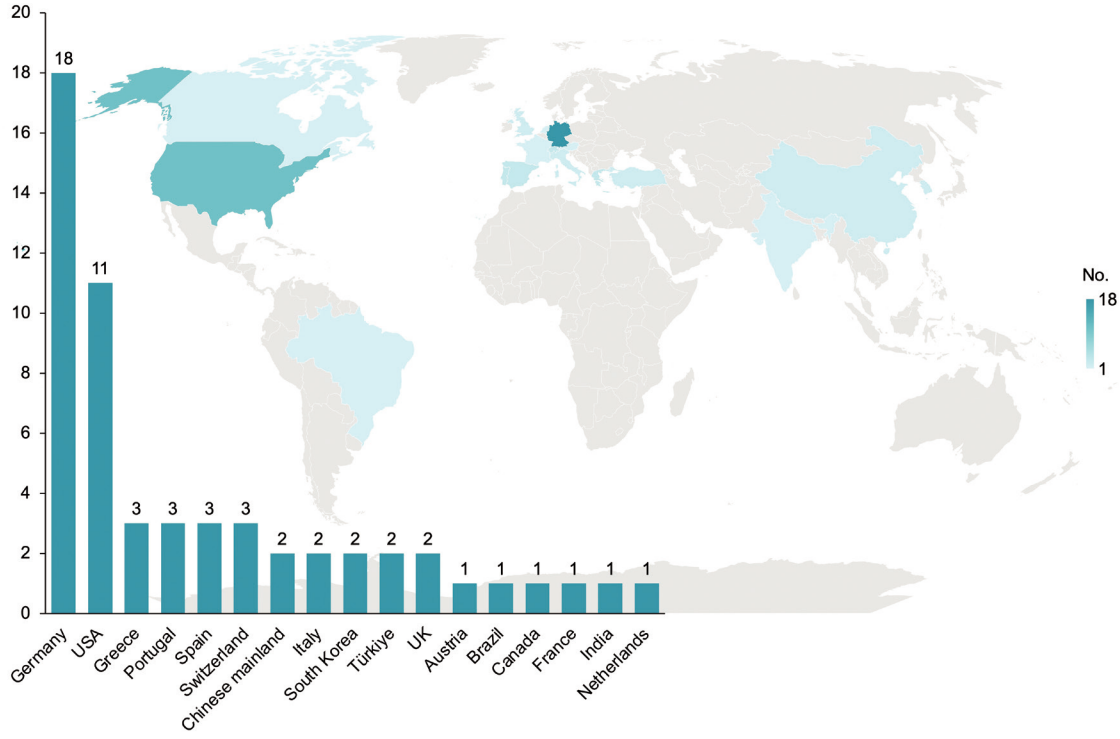

Figure 2. Combined country and region distribution of the panelists and steering committee.

No, number of contributors.

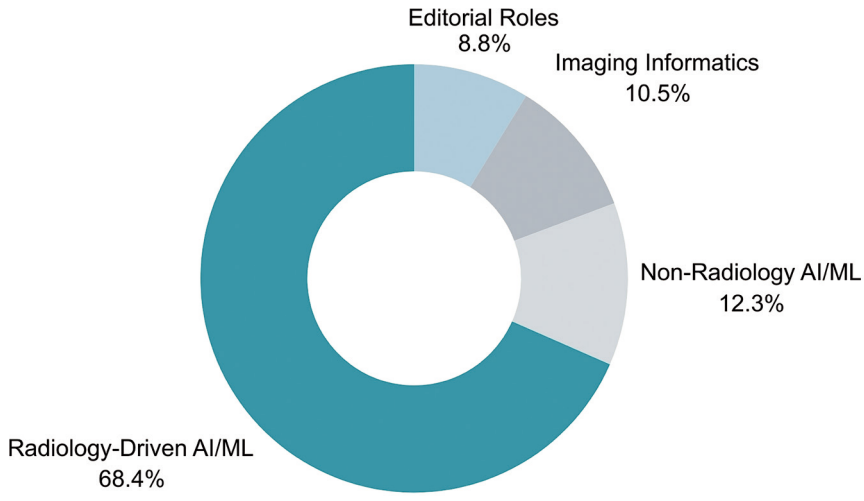

Figure 3. Committee-assigned expertise distribution across the REFINE Delphi panelists and steering committee members.

REFINE, REporting checklist for FoundatIon and large laNguagE models; AI, artificial intelligence; ML, machine learning.

their research. Figure 5 illustrates the main functionalities of the web version of the REFINE.

Discussion

Principal findings

In this study, we developed the REFINE, a consensus-based reporting guideline designed to address the opacity and heterogeneity of FMs and LLMs in medical research. Unlike general AI reporting guidelines, the REFINE explicitly targets sources of variability and risks unique to generative AI, spanning model specification, prompt design, stochasticity control, dataset integrity, output evaluation, and implementation. By grounding the checklist in a formal international Delphi consensus process, the REFINE provides a pragmatic standard to improve the quality, consistency, and reproducibility of this rapidly evolving field. Although the consensus panel included strong representation from imaging-related disciplines, the resulting checklist items, particularly those governing prompt engineering, stochasticity control, and dataset contamination, address fundamental properties of FMs and LLMs that apply to text-only, multimodal, and imaging workflows alike.

Relation to existing guidelines

The REFINE is designed to complement established EQUATOR-aligned guidelines. Frameworks such as CLAIM,[29,30] CONSORT-AI,[31] TRIPOD-AI,[32] and STARD-AI[33] provide a robust foundation for study design, participant selection, reference standards, and performance metrics but were developed before the widespread adoption of generative AI. Consequently, they offer limited coverage of several characteristics specific to FMs and LLMs, such as stochasticity and prompt engineering.

Recent efforts have emerged to address this reporting gap.[17-19,34-37] The TRIPOD-LLM framework extends TRIPOD-AI using a modular checklist to cover model development and evaluation, specifically within the context of diagnostic and prognostic prediction models.[18] Similarly, MI-CLEAR-LLM establishes minimum reporting items for accuracy

Table 1. Standardized terminology used in the REFINE

Term	Definition
Training	The process of optimizing a model's parameters using data and an objective function to minimize a defined loss metric through iterative gradient updates.
Pretraining	Training performed on large-scale, general, or weakly supervised datasets, often via self-supervised learning, to develop broad representations and foundational capabilities.
Post-training	The subsequent optimization phase that includes supervised fine-tuning, instruction tuning, task-specific or domain-specific adaptation, and alignment methods such as reinforcement learning from human feedback (RLHF), reinforcement learning from artificial intelligence feedback (RLAIF), and direct preference optimization (DPO). This phase refines the pretrained model's behavior toward human intent, safety, or specialized performance.
Fine-tuning	A targeted form of post-training focused on adapting model weights for a particular task, dataset, or domain to achieve an explicit objective such as classification accuracy or instruction adherence.
Inference-time adaptation	Adaptation without weight updates, achieved through context conditioning or external mechanisms such as prompting, retrieval-augmented generation, tool integration, or dynamic hyperparameter selection.
Alignment	Steering model behavior toward human intent or domain goals, achieved through weight-updating post-training methods such as RLHF, RLAIF, or DPO or through inference-time methods such as prompting or retrieval.
Testing	Final, strictly unseen hold-out evaluation performed once, with no design choices informed by test feedback.
Validation	Any confirmation process that does not refer to a held-out data partition (i.e., not validation or testing splits). Throughout this checklist, "validation" refers to verifying the correctness or adequacy of procedures and is not used to denote a machine learning validation dataset due to its ambiguity.

REFINE, **RE**porting checklist for **F**oundat**I**on and large la**N**guag**E** models.

reports in healthcare, with a specific focus on handling stochasticity, prompt syntax transparency, and model access modes.[17,37] To accommodate varying levels of technical depth, the DEAL checklist introduces dual pathways, one for advanced model development and another for off-the-shelf applications.[19]

Other initiatives target specific use cases or ethical dimensions. The CHART statement focuses on studies evaluating chatbot health advice, emphasizing query strategies and prompt engineering for clinical advice summarization;[34] CANGARU addresses the ethical use and disclosure of generative AI tools within the academic writing and publishing process itself.[36] Additionally, CRAFT-MD pro-

Section	Item	Keep as is	Keep with Modification	Remove	Unsure	Consensus to Keep	Origin of Statistics
1. Model Specification	1.1 Model name, vendor/developer, version/identifier, release date, and training/knowledge cutoff date	79.6	20.4	0.0	0.0	100.0	Round 1
	1.2 Model architecture and key characteristics	83.3	7.4	7.4	1.9	92.5	Round 2 (re-voted)
	1.3 Model pretraining, post-training, and inference-time adaptation strategy	85.2	11.1	1.9	1.9	98.1	Round 2 (re-voted)
	1.4 Modality support details (input and output) and limitations	83.3	16.7	0.0	0.0	100.0	Round 2 (re-voted)
	1.5 Language capabilities	79.6	9.3	5.6	5.6	94.1	Round 2 (re-voted)
	1.6 Model access	77.8	20.4	1.9	0.0	98.1	Round 1
	1.7 Sharing of code, data, and model artifacts	81.5	18.5	0.0	0.0	100.0	Round 2 (proposed)
	1.8 Computational requirements	75.9	20.4	1.9	1.9	98.1	Round 1
2. Prompt Design	2.1 Prompt engineering protocol with versioning	90.7	3.7	1.9	3.7	98.0	Round 2 (proposed)
	2.2 Prompting strategy, format, and length	77.8	19.45	1.85	0.95	98.1	Round 1*
	2.3 Prompt modality, language, technical input specification, and full content	94.4	3.7	1.9	0.0	98.1	Round 2 (proposed)
	2.4 Integration of relevant patient clinical context	83.3	11.1	1.9	3.7	98.0	Round 1
	2.5 Interaction style and session memory policy	70.4	16.7	9.3	3.7	90.4	Round 1
	2.6 Output handling	81.5	11.1	3.7	3.7	96.2	Round 2 (proposed)
3. Stochasticity Control	3.1 Generation parameters	76.9	15.8	1.9	5.6	98.0	Round 1*
	3.2 Prompt operator characteristics and number of prompt attempts	70.4	18.5	3.7	7.4	96.0	Round 1
	3.3 Output selection	75.9	13.0	1.9	9.3	97.9	Round 1
4. Dataset Integrity	4.1 Dataset name, version, access type, source citation, license, and compliance statement	83.3	16.7	0.0	0.0	100.0	Round 1
	4.2 Dataset origin	83.3	16.7	0.0	0.0	100.0	Round 2 (proposed)
	4.3 Ethics and consent statements	92.6	5.6	1.9	0.0	98.1	Round 1
	4.4 Prior dataset usage and publication date	74.1	14.8	5.6	5.6	94.1	Round 1
	4.5 Dataset composition and data synthesis details	74.1	24.1	1.9	0.0	98.1	Round 1
	4.6 Sample characteristics and representational bias analysis	85.2	14.8	0.0	0.0	100.0	Round 1
	4.7 Reference standard and annotator qualifications	83.3	14.8	0.0	1.9	100.0	Round 1
	4.8 Preprocessing and data pairing/registration	87.0	13.0	0.0	0.0	100.0	Round 1
	4.9 Missing data extent, mechanism, and handling	92.6	5.6	1.9	0.0	98.1	Round 1
	4.10 Separation of training, fine-tuning, internal testing, and external testing datasets	70.4	27.8	0.0	1.9	100.0	Round 1
5. Output Evaluation	5.1 Output evaluation method and performance metrics	78.7	21.3	0.0	0.0	100.0	Round 1*
	5.2 Human evaluator characteristics and reliability analysis	90.7	9.3	0.0	0.0	100.0	Round 1
	5.3 Statistical analysis of evaluation results	94.4	5.6	0.0	0.0	100.0	Round 2 (proposed)
	5.4 Subgroup performance and output bias assessment	79.7	11.1	3.7	5.6	96.1	Round 1*
	5.5 Failure analysis and error metrics	83.3	13.0	1.9	1.9	98.1	Round 2 (proposed)
	5.6 Output stochasticity and reproducibility constraints	81.5	13.0	1.9	3.7	98.0	Round 1
	5.7 Performance effects of different prompt strategies and revisions	66.7	13.0	11.1	9.3	87.8	Round 1
	5.8 Model version comparisons and temporal performance variation	64.8	22.2	7.4	5.6	92.2	Round 1
	5.9 Methods for explainability and interpretability of model outputs	81.5	11.1	3.7	3.7	96.2	Round 1
	5.10 Comparison with clinically relevant benchmarks	77.8	14.8	3.7	3.7	96.2	Round 1
6. Implementation	6.1 Declared intended application and scope of use	72.2	22.2	1.9	3.7	98.0	Round 1
	6.2 Clinical workflow integration	79.6	13.0	1.9	5.6	98.0	Round 2 (proposed)
	6.3 Measured clinical utility or added value	90.7	7.4	1.9	0.0	98.1	Round 1
	6.4 Model limitations, explicit clinical non-use cases, and potential misuse considerations	59.3	22.2	13.0	5.6	86.2	Round 1
	6.5 Safety testing and monitoring protocols	72.2	20.4	3.7	3.7	96.2	Round 1
	6.6 Data security and privacy safeguards	85.2	7.4	3.7	3.7	96.2	Round 2 (proposed)**
	6.7 Governance, auditability, and oversight	85.2	7.4	3.7	3.7	96.2	Round 2 (proposed)**

Figure 4. Consensus voting results for the finalized REFINE items. Bars indicate the percentage of panelists selecting each response option for each item across six sections. The rightmost column summarizes consensus to keep using a heat map, and the final column indicates the Delphi round from which the statistics were derived. Percentages may not sum to 100 due to rounding.

**Average of merged items; ** Split from a parent item. REFINE, REporting checklist for FoundatIon and large laNguagE models.*

Table 2. The REFINE checklist

Section	Item	Reported: Yes	Partial	No	N/A	Location
1. Model specification	1.1 Model name, vendor/developer, version/identifier, release date, and training/knowledge cutoff date	☐	☐	☐	☐	
	1.2 Model architecture and key characteristics	☐	☐	☐	☐	
	1.3 Model pretraining, post-training, and inference-time adaptation strategy	☐	☐	☐	☐	
	1.4 Modality support details (input and output) and limitations	☐	☐	☐	☐	
	1.5 Language capabilities	☐	☐	☐	☐	
	1.6 Model access	☐	☐	☐	☐	
	1.7 Sharing of code, data, and model artifacts	☐	☐	☐	☐	
	1.8 Computational requirements	☐	☐	☐	☐	
2. Prompt design	2.1 Prompt engineering protocol with versioning	☐	☐	☐	☐	
	2.2 Prompting strategy, format, and length	☐	☐	☐	☐	
	2.3 Prompt modality, language, technical input specification, and full content	☐	☐	☐	☐	
	2.4 Integration of relevant patient clinical context	☐	☐	☐	☐	
	2.5 Interaction style and session memory policy	☐	☐	☐	☐	
	2.6 Output handling	☐	☐	☐	☐	
3. Stochasticity control	3.1 Generation parameters	☐	☐	☐	☐	
	3.2 Prompt operator characteristics and number of prompt attempts	☐	☐	☐	☐	
	3.3 Output selection	☐	☐	☐	☐	
4. Dataset integrity	4.1 Dataset name, version, access type, source citation, license, and compliance statement	☐	☐	☐	☐	
	4.2 Dataset origin	☐	☐	☐	☐	
	4.3 Ethics and consent statements	☐	☐	☐	☐	
	4.4 Prior dataset usage and publication date	☐	☐	☐	☐	
	4.5 Dataset composition and data synthesis details	☐	☐	☐	☐	
	4.6 Sample characteristics and representational bias analysis	☐	☐	☐	☐	
	4.7 Reference standard and annotator qualifications	☐	☐	☐	☐	
	4.8 Preprocessing and data pairing/registration	☐	☐	☐	☐	
	4.9 Missing data extent, mechanism, and handling	☐	☐	☐	☐	
	4.10 Separation of training, fine-tuning, internal testing, and external testing datasets	☐	☐	☐	☐	
5. Output evaluation	5.1 Output evaluation method and performance metrics	☐	☐	☐	☐	
	5.2 Human evaluator characteristics and reliability analysis	☐	☐	☐	☐	
	5.3 Statistical analysis of evaluation results	☐	☐	☐	☐	
	5.4 Subgroup performance and output bias assessment	☐	☐	☐	☐	
	5.5 Failure analysis and error metrics	☐	☐	☐	☐	
	5.6 Output stochasticity and reproducibility constraints	☐	☐	☐	☐	
	5.7 Performance effects of different prompt strategies and revisions	☐	☐	☐	☐	
	5.8 Model version comparisons and temporal performance variation	☐	☐	☐	☐	
	5.9 Methods for explainability and interpretability of model outputs	☐	☐	☐	☐	
	5.10 Comparison with clinically relevant benchmarks	☐	☐	☐	☐	
6. Implementation	6.1 Declared intended application and scope of use	☐	☐	☐	☐	
	6.2 Clinical workflow integration	☐	☐	☐	☐	
	6.3 Measured clinical utility or added value	☐	☐	☐	☐	
	6.4 Model limitations, explicit clinical non-use cases, and potential misuse considerations	☐	☐	☐	☐	
	6.5 Safety testing and monitoring protocols	☐	☐	☐	☐	
	6.6 Data security and privacy safeguards	☐	☐	☐	☐	
	6.7 Governance, auditability, and oversight	☐	☐	☐	☐	

REFINE, **RE**porting checklist for **F**oundat**I**on and large la**N**guag**E** models; N/A; not applicable.

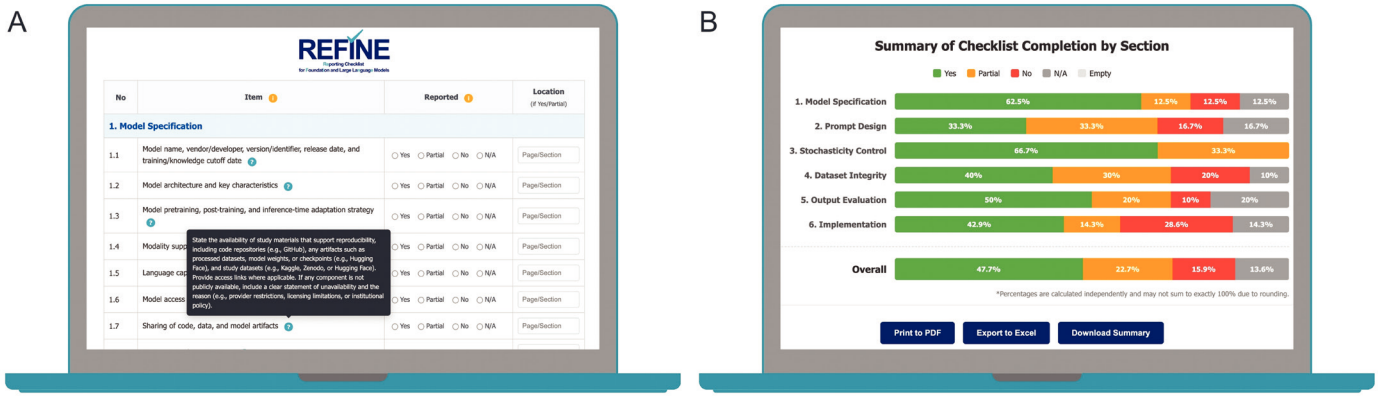

Figure 5. Web-based interface of the REFINE checklist. Panel A shows a portion of the interactive checklist—hovering over the information icon next to each item reveals detailed reporting instructions. Panel B shows the automated summary of checklist completion by section, with options to print the checklist to PDF, export the data to Excel, and download the summary statistics image.

vides a framework specifically for evaluating conversational reasoning through simulated doctor–patient interactions rather than a general reporting structure for study methodology.[35]

The REFINE distinguishes itself within this ecosystem by integrating technical reproducibility with broader implementation governance. Although guidelines such as MI-CLEAR-LLM focus on the details of accuracy testing (e.g., temperature settings, prompt syntax) to some extent, the REFINE expands these requirements and extends them across the full study lifecycle, mandating reporting on dataset integrity (e.g., contamination risks, representational bias) and clinical implementation (e.g., workflow integration, failure analysis, and safety protocols). Thus, the REFINE serves as a comprehensive standard for documenting both the generative parameters and the clinical reliability of FM and LLM studies.

The REFINE is also intended to be used alongside other AI reporting tools. For example, a randomized trial involving an LLM would report the trial design using CONSORT-AI and the model methodology using the REFINE.

Contributions of the REFINE

The REFINE introduces critical reporting requirements that address the non-deterministic nature of generative AI. First, it mandates detailed reporting of model specifications. Unlike traditional algorithms, models with similar names may differ considerably due to access configuration, quantization, tooling, and safety alignment layers, all of which determine validity and generalizability.[38-43] Second, the REFINE requires explicit documentation of prompt engineering protocols with the same rigor as code in deterministic algorithms, including the specific context provided. Third, it enforces detailed reporting of generation parameters (e.g., temperature, top-p), which can significantly reshape output distributions and are critical for model performance and reproducibility.[13,24,44,45] Without these, identical models may produce divergent outputs, rendering a study irreproducible. Fourth, the REFINE addresses dataset integrity by assessing the risk of contamination (i.e., overlap between evaluation datasets and the model's pretraining corpus), which is a major challenge for fairly evaluating FM and LLM performance.[46,47] Fifth, the REFINE emphasizes structured reporting of interaction style, session memory, tool use, retrieval-augmented generation, and multimodal integration, which are central to modern FM and LLM applications. Finally, the REFINE incorporates implementation-focused items, requiring authors to report monitoring for misuse and failure modes specific to clinical workflows.

Practical use and implementation

The REFINE serves as a comprehensive, practical tool for multiple stakeholders. For authors, the core checklist acts as a prospective design aid to ensure key elements are considered during study planning, whereas detailed item instructions support manuscript preparation. For reviewers and editors, the REFINE can serve as a structured appraisal tool to systematically evaluate methodological transparency, reducing reliance on individual familiarity with rapidly evolving technical details. It can also help identify specific gaps that limit interpretability or reproducibility.

The REFINE has the potential to be adopted and reinforced at the level of journals, conferences, and professional societies. We propose that journals integrate the REFINE into their author instructions and editorial policies to normalize the use of these standards. Endorsement by major bodies or societies may facilitate broader adoption.

Strengths and limitations

The REFINE has several notable strengths. First, it was developed by an international and multidisciplinary panel, which supports its applicability across settings. Second, the checklist was developed through a predefined and transparent Delphi process with explicit consensus thresholds and decision rules, thereby reducing the risk of bias. Third, the availability of a user-friendly online platform further facilitates practical and consistent use. Fourth, the REFINE is applicable across diverse study designs; the inclusion of an "N/A" option functions as a deliberate filtering mechanism, allowing investigators to exclude non-applicable items without penalizing overall checklist completion.

The REFINE also has several limitations. Although the panel was international and multidisciplinary, its composition may still introduce bias, including a predominance of imaging experts and an underrepresentation of certain geographies, specialties, and stakeholder groups. Consequently, some domain-specific reporting needs, particularly those outside imaging-intensive disciplines or resource-rich healthcare contexts, may not be fully captured. Furthermore, although the checklist was developed via expert consensus, formal pilot testing with external users to validate usability was not conducted before release. In addition, the modified Delphi process, though systematic, remains dependent on subjective judgments. Finally, the REFINE was developed in the context of

Table 3. The REFINE reporting instructions

Section	Items and reporting instructions
1. Model specification	**1.1 Model name, vendor/developer, version/identifier, release date, and training/knowledge cutoff date.** Report the full model identity. Specify the model name (e.g., GPT-5), vendor or developer (e.g., OpenAI or Stanford University), version or application programming interface (API) identifier (e.g., 5 or GPT-5-2025-08-07), and model release date. If an official release date is unavailable, report the date accessed and cite the model card or changelog version. Also report the training or knowledge cutoff date (if provided), which defines the latest point up to which the model was trained.
	1.2 Model architecture and key characteristics. Report the model's foundational architecture (e.g., transformer or state-space models) and any key design features (e.g., mixture-of-experts or diffusion-based models). Include parameter count if available. For multimodal models, specify the image encoder architecture (e.g., ResNet or vision transformer) and the fusion strategy (e.g., cross-attention) if available.
	1.3 Model pretraining, post-training, and inference-time adaptation strategy. Describe how the model was developed or adapted for the study task. Indicate whether the study involved training a new foundation model (pretraining from scratch or continued pretraining) or post-training an existing one. For training, summarize the pretraining corpus and objectives. For post-training, specify whether it involved tuning or alignment and whether model weights were updated. If weights were updated, state the method (e.g., supervised fine-tuning or reinforcement learning) and the technique (e.g., full fine-tuning or parameter-efficient fine-tuning), including key details such as dataset size, steps/epochs, batch size, and optimizer. If no weight updates were performed, describe the inference-time adaptation strategy (e.g., prompting or retrieval-augmented generation), and report any external tools or APIs used at inference. Indicate whether clinical data were included, and clearly specify the stage(s) at which they were used (pretraining, post-training, or inference-time adaptation).
	1.4 Modality support details (input and output) and limitations. Report the input and output modalities supported by the model (e.g., text, image, audio, and video). Describe any technical limitations, such as maximum context length (e.g., up to 128k tokens), image resolution constraints (e.g., ≤1,024 × 1,024 pixels), patch size, or input-pipeline restrictions (e.g., DICOM files require prior conversion to image tensors; only 2D images are accepted).
	1.5 Language capabilities. Report the languages in which the model has been evaluated or explicitly tested, and specify any known limitations or domain specialization (e.g., medical English, radiology-specific Turkish, or German for lay explanations). If applicable, indicate whether performance across languages was assessed or remains unverified.
	1.6 Model access. Describe how the model was accessed during the study. Specify whether access was through a graphical user interface (e.g., chat platform) or an API. For API access, indicate whether it was locally hosted, securely hosted (e.g., an enterprise cloud), or publicly hosted.
	1.7 Sharing of code, data, and model artifacts. State the availability of study materials that support reproducibility, including code repositories (e.g., GitHub), any artifacts such as processed datasets, model weights, or checkpoints (e.g., Hugging Face), and study datasets (e.g., Kaggle, Zenodo, or Hugging Face). Provide access links where applicable. If any component is not publicly available, include a clear statement of unavailability and the reason (e.g., provider restrictions, licensing limitations, or institutional policy).
	1.8 Computational requirements. Report the computational resources required for model development and use. Specify hardware and resource needs for different stages, including training, fine-tuning, alignment, and inference (e.g., graphics processing unit/tensor processing unit type, number of compute nodes, memory requirements, runtime, or cloud compute specifications).
2. Prompt design	**2.1 Prompt engineering protocol with versioning.** Report the protocol used for prompt engineering and the development process [e.g., iterative testing of prompt variants with predefined success metrics, human-in-the-loop review with inter-rater checks, or automated prompt optimization (e.g., DSPy)]. Describe contributors involved (e.g., domain experts or engineers), data partitioning used during prompt development, and any automated tools or optimization frameworks applied. Provide a version history of prompts, summarizing major changes across iterations.
	2.2 Prompting strategy, format, and length. Describe how prompts were constructed and used. Specify the prompting strategy (e.g., zero-shot with task instructions, few-shot using exemplar inputs, or chain-of-thought prompting), the prompt format (e.g., structured templates with placeholders or open-ended queries for text prompts; a single image, a multi-image set, or region annotations/bounding boxes for image prompts), and prompt length (e.g., short directive prompts vs. long multi-context prompts).
	2.3 Prompt modality, language, technical input specification, and full content. Specify the prompt modality (e.g., text, image, or audio), type of input(s) (e.g., chest X-ray images, discharge summary text, or pathology reports), and the language (if text prompts are used or images contain text). Include relevant technical characteristics such as resolution, encoding, or preprocessing applied before input conversion. Specify whether and how prior studies or longitudinal data were included in the prompt for comparison. Provide the full prompt content used in the study (e.g., verbatim text for text prompts or representative examples for non-text prompts).
	2.4 Integration of relevant patient clinical context. Describe how patient-specific clinical information was incorporated into prompts. Specify what types of context were included (e.g., age, sex, key comorbidities, prior treatments, and relevant medical history), how this information was selected (e.g., based on predefined criteria, guideline-driven relevance, or expert curation), and how it was standardized (e.g., ICD codes for diagnoses, SNOMED CT for procedures, RxNorm for medications, and BI-/LI-/PI-RADS for imaging). Also state the source of the clinical information (e.g., electronic health records, radiology reports, or patient summaries).
	2.5 Interaction style and session memory policy. Specify how users interacted with the model. Indicate whether the workflow used a single-turn interaction (independent one-shot queries) or a multi-turn conversation in which previous messages influence later responses. Also report the memory policy, whether prior context was retained across messages in the same conversation and reset after closing the session (i.e., session memory policy) or whether prior context was retained across different conversations (i.e., persistent memory policy), to clarify how much conversational history shaped the model's outputs.

Table 3. Continued

Section	Items and reporting instructions
2. Prompt design	**2.6 Output handling.** Report how model outputs were controlled and managed after generation. Describe the output format used (e.g., structured JSON, free text, or tabular form). Specify the level of control applied: i, no control; ii, control in prompt only; iii, control during generation (e.g., guided generation or JSON mode with schema validation); or iv, control after generation (e.g., checking for information completeness, validating against a ground truth, or validating against a schema). Describe how constraints or output schemas were enforced and any further validation schemes applied (e.g., clinical plausibility review).
3. Stochasticity control	**3.1 Generation parameters.** Report all model generation settings used during output generation for all modalities (e.g., text, image, and others). For text generation, specify parameters such as temperature, top-k/top-p sampling, maximum output tokens, repetition or frequency penalties, and any random seed used. For image generation, include the number of inference steps, scheduler type, output resolution, and guidance scale. **3.2 Prompt operator characteristics and number of prompt attempts.** Specify who interacted with the model. Report these operators' roles (e.g., clinician, researcher, attending radiologist, resident, or technologist) and experience levels (e.g., years of practice, AI familiarity, or training status). Also specify how many prompt attempts were made. **3.3 Output selection.** Describe how the final model output(s) were selected. If multiple generations were sampled, detail the selection criteria. Specify whether final outputs were expert-reviewed (e.g., radiologist selected best response), randomly selected (e.g., first output used without filtering), consensus-based (e.g., agreement among multiple reviewers), algorithmic (e.g., ranking by confidence score), or AI-automated (e.g., pipeline-selected output). Report tie-break rules and any rejection filters.
4. Dataset integrity	**4.1 Dataset name, version, access type, source citation, license, and compliance statement.** Clearly report the dataset used to ensure transparency and traceability. Include the dataset name and version (e.g., MIMIC-CXR v2.0, LIDC-IDRI, or BraTS 2023), access type (public, restricted, or private), and citation of the data source. For private datasets, specify the institution or repository name and how the data were accessed (e.g., through departmental electronic health records, institutional PACS, or a trial repository). Also include license details (e.g., PhysioNet Credentialed Health Data License; RSNA data agreement) and a statement confirming compliance with data use agreements or institutional approvals. **4.2 Dataset origin.** Specify whether the dataset was collected from a single site or multiple centers. For multi-center datasets, indicate the number and type of institutions (e.g., academic medical centers or community hospitals) and whether the data were international or regional. **4.3 Ethics and consent statements.** Specify institutional review board or ethics committee approval status, including the approval number if applicable. If informed consent was waived, provide a brief justification (e.g., retrospective anonymized data). **4.4 Prior dataset usage and publication date.** Report any prior usage of the dataset that could affect the evaluation of independence, particularly prior work by the authors or institution. For widely used public datasets, a general acknowledgment of their established use is sufficient. Also state the dataset's publication or public release date, as datasets made public before a model's pretraining cutoff may carry a risk of contamination. If usage details are limited by proprietary restrictions, explicitly state the limitations and their source. **4.5 Dataset composition and data synthesis details.** Report the composition of the dataset, specifying whether it includes real clinical data, synthetic data, or a combination of both. For public datasets, citing the original source and providing a brief description is sufficient. For private or newly published datasets, disclose the data generation method (e.g., LLM-based report synthesis, diffusion model for MR image synthesis, or generative adversarial network-based augmentation of CT scans) and the proportion of synthetic data in the dataset. **4.6 Sample characteristics and representational bias analysis.** Report key characteristics of the dataset population to assess fairness and generalizability. Include demographics (e.g., age distribution and sex balance), clinical characteristics (e.g., disease types and severity levels), and data composition (e.g., imaging modality, number of classes, and case distribution). Also evaluate representational bias by analyzing subgroup coverage (e.g., underrepresentation of certain age groups, sex imbalance, or limited geographic diversity). **4.7 Reference standard and annotator qualifications.** Define how the reference standard in the dataset was established. Specify the type of reference standard used (e.g., pathology-confirmed diagnosis, radiology report, clinical follow-up, or expert panel consensus) and describe the annotator qualifications (e.g., board-certified radiologist with 10 years of experience, pathology fellow, or multidisciplinary tumor board). Include the number of annotators and any disagreement-resolution process. **4.8 Preprocessing and data pairing/registration.** Describe all preprocessing steps applied to the dataset before model use. For text data, specify methods for de-identification/anonymization (e.g., removal of protected health information or natural language processing-based redaction). For imaging data, report pairing or registration procedures (e.g., linking reports to imaging studies, spatial alignment across sequences, or longitudinal time points). Report specific processing pipelines such as DICOM window/level settings, normalization techniques, and multi-sequence handling. Also mention any filtering, resizing, artifact removal, or other normalization procedures. **4.9 Missing data extent, mechanism, and handling.** Report how missing data were assessed and managed. Specify the extent of missing data (e.g., percentage of missing values per variable), the mechanism if known (e.g., missing completely at random, missing at random, or missing not at random), and the handling strategy used (e.g., exclusion of incomplete cases, mean/median imputation, model-based imputation, or no imputation). Provide a brief rationale for the chosen method to support transparency and reproducibility.

Table 3. Continued

Section	Items and reporting instructions
4. Dataset integrity	**4.10 Separation of training, fine-tuning, internal testing, and external testing datasets.** Report how datasets were separated to prevent information leakage. Clearly describe the distinction between training data, fine-tuning data, internal testing data (i.e., held-out cases from the same institution), and external testing data (i.e., from a different institution). State how independence was ensured (e.g., separation by patient ID, study date, or site). Confirm that no internal or external test data were used during training, fine-tuning, or prompt optimization.
5. Output evaluation	**5.1 Output evaluation method and performance metrics.** Report the methods used to evaluate the model and list all relevant performance metrics. State their appropriateness for the task(s) of interest. Specify whether the evaluation was based on human review (e.g., Likert ratings, readability scores, or clinical usefulness ratings), task-performance metrics (e.g., AUROC, F1-score, sensitivity, specificity, or calibration measures), text or semantic similarity metrics (e.g., BLEU, ROUGE, or BERTScore), or model-based evaluation (e.g., LLM-as-a-judge scoring). **5.2 Human evaluator characteristics and reliability analysis.** Report the characteristics of human evaluators involved in the assessment of model outputs, including the number of evaluators, their role or specialty (e.g., radiologist, clinician, or domain expert), experience level, and any formal training provided for the evaluation task. Report the methods used to measure evaluator consistency, including inter-rater or intra-rater reliability statistics (e.g., Cohen's kappa, Fleiss' kappa, or intraclass correlation coefficient). **5.3 Statistical analysis of evaluation results.** Report the statistical methods used to analyze model evaluation outcomes. Specify any hypothesis testing or alternative approaches, the interval estimates of uncertainty used (e.g., confidence intervals, credible intervals, or bootstrap intervals), and effect size measures where applicable. Describe the criteria used to interpret statistical evidence (e.g., significance thresholds, Bayesian decision rules, or equivalence margins). **5.4 Subgroup performance and output bias assessment.** Report model performance across predefined subgroups relevant to the clinical task, such as age groups, sex, disease categories, imaging modality, institution, or geographic region. Describe any observed performance differences between subgroups to examine potential output bias. **5.5 Failure analysis and error metrics.** Report how model errors were identified, reviewed, and categorized. Describe the procedure used to detect and classify failures (e.g., hallucination, reasoning error, bias-related error, factual inaccuracy, or formatting issue). Provide representative examples of failure types. Report key error metrics (e.g., hallucination rate, factual inaccuracy rate, or omission rate) and include summaries of error distribution where available. **5.6 Output stochasticity and reproducibility constraints.** Report the results of generative variability assessment based on repeated generations for identical prompts and randomness settings. Provide examples of output differences. Based on the analysis, state any factors limiting exact reproducibility, such as inherent stochastic model behavior, lack of seed control, proprietary or closed-source model access, updates to model versions over time, or restricted access environments where full configuration cannot be disclosed. **5.7 Performance effects of different prompt strategies and revisions.** Report how different prompting strategies affected model performance, including comparisons across approaches such as zero-shot, few-shot, chain-of-thought prompting, or prompting using different languages. Report the impact of any prompt revisions (e.g., initial prompt version vs. revised version) on output quality or task performance. **5.8 Model version comparisons and temporal performance variation.** Report comparisons of model performance across different versions of the model (e.g., Llama-3.1-8b v1.0 or Llama-3.1-8b v1.1) or across different release dates of the same version, if such evaluations were performed or relevant to the study period. Describe how performance changed over time or between versions, and specify the conditions used for comparison. **5.9 Methods for explainability and interpretability of model outputs.** Report the methods used to interpret or explain model outputs and describe how they were applied, as applicable to the model type and study context. Describe the scope and known limitations of the chosen explainability approach in the study context. **5.10 Comparison with clinically relevant benchmarks.** Report comparisons between model performance and appropriate clinical benchmarks. These benchmarks may include established clinical standard-of-care (e.g., expert human performance or clinical guidelines). Comparisons to specialist task-specific artificial intelligence models should be reported as technical benchmarks, distinct from clinical reference standards. Describe the comparison framework and reference standards used.
6. Implementation	**6.1 Declared intended application and scope of use.** Report the intended purpose of the model. Describe the target application, such as diagnostic decision support, medical question answering, report generation, workflow triage, patient communication, or educational support. **6.2 Clinical workflow integration.** Report whether, and if applicable, how and where the model was integrated into the clinical workflow. Specify the integration mode (e.g., embedded in PACS or RIS, decision-support dashboard, web-based interface, or standalone software) and the interaction point in the workflow (e.g., prereading triage, concurrent reading, post-report review, or quality assurance). State the intended user role (e.g., radiologist, resident, or technologist) and whether patient-facing interaction was involved. **6.3 Measured clinical utility or added value.** Report measures of clinical utility or added value obtained from model use. Describe how the model contributed to outcomes such as improvement in diagnostic accuracy, reduction in reporting or decision time, enhancement of workflow efficiency, increased clinician confidence, or improved patient understanding. **6.4 Model limitations, explicit clinical non-use cases, and potential misuse considerations.** Report known limitations of the model and clearly state clinical scenarios where it should not be used. Describe explicit non-use cases (e.g., high-risk decision-making without human oversight, unsupported imaging modalities, or unsupported patient populations) and identify foreseeable risks of misuse. Include considerations related to potential patient harm, safety concerns, or health system risks.

Table 3. Continued

Section	Items and reporting instructions
6. Implementation	**6.5 Safety testing and monitoring protocols.** Report procedures used to identify and manage harmful, clinically unsafe, or medically inaccurate outputs. Describe any safety testing performed before deployment (e.g., screening for harmful recommendations or toxic responses and sandboxing) and monitoring protocols used during model interaction (e.g., automated safety filters or human review of unsafe outputs).
	6.6 Data security and privacy safeguards. Report measures used to protect data security and patient privacy during model use if patient or sensitive data were processed. Describe how prompt and user data were handled, including storage policies, duration of retention, and whether data were reused for model improvement. Specify data routing and regional processing locations if applicable (e.g., EU vs. US data centers) and report safeguards for live handling of protected health information, such as de-identification, access control, and encryption.
	6.7 Governance, auditability, and oversight. Report governance measures and institutional oversight applied to the use of the model in the study, distinct from the initial ethics approval. Specify procedures for auditability (e.g., logging of model interactions or traceability of outputs) and formal oversight, especially when using external or API-based models (e.g., use of secure institutional accounts vs. public APIs).

REFINE, **RE**porting checklist for **F**oundat**I**on and large la**N**guag**E** models.

rapidly evolving FM and LLM technologies, regulatory expectations, and clinical use cases. The checklist, therefore, reflects the best available knowledge but requires adaptation and updates as model capabilities evolve.

Future directions and planned updates

We plan to update the REFINE through a formal re-evaluation of its items every 2 years, guided by feedback from users and the community, developments in related reporting standards, and emerging evidence on FM and LLM deployment in healthcare. In parallel, future work may also explore domain-specific extensions or modular add-ons such as radiology-focused variants, imaging-intensive implementations, text-only clinical documentation modules, and decision-support modules while preserving a common core.

An additional priority is to evaluate the uptake, usability, and impact of the REFINE in practice. This may include surveys or qualitative studies of authors, reviewers, and editors; bibliometric analyses of reporting quality before and after journal endorsement; and targeted audits of FM and LLM studies using the REFINE. These evaluations will help identify challenging items, clarify where further guidance is needed, and determine how the REFINE can best support transparent and high-quality reporting as the field evolves.

Final remarks

The integration of FM and LLM into medicine demands reporting standards that match their complexity, risks, and clinical implications. Without rigorous documentation, evidence generated from these systems will remain difficult to trust and reproduce. The REFINE directly addresses this gap by providing a consensus-built framework that clarifies what must be documented. Its adoption offers a practical foundation for transparent, reproducible, and ultimately trustworthy medical AI research.

Acknowledgements

Language of this manuscript was checked and improved by generative AI (ChatGPT-5 and 5.2; Gemini 2.5 and 3 Pro). The authors conducted strict supervision when using these tools.

Funding

This study received no specific funding.

Footnotes

Conflict of interest disclosure

T. Akinci D'Antonoli serves as Section Editor for Diagnostic and Interventional Radiology. She had no involvement in the peer-review of this article and had no access to information regarding its peer-review. **A. Chaudhari** receives research support from GE Healthcare, Philips, Microsoft, Amazon, Google, NVIDIA, and Stability; provides consulting services to Patient Square Capital, Chondrometrics GmbH, Elucid Bioimaging, and Cognita Imaging; is a co-founder of Cognita Imaging; and holds equity interests in Subtle Medical, LVIS Corp, Brain Key, and Radiology Partners. **B. Khosravi** serves as Associate Editor of Radiology: Artificial Intelligence. **C.E. Kahn Jr.** serves as Editor of Radiology: Artificial Intelligence. **D.M. Koh** provides consultancy to GE Healthcare and GlaxoSmithKline (GSK) and maintains research collaborations with Siemens Healthineers, QED, and Mint Medical. **F. Kitamura** is a consultant for Bunkerhill Health, GE Healthcare, and MD.ai; a speaker for Sharing Progress in Cancer Care; holds leadership roles as Early Career Consultant to the Editor of Radiology, Associate Editor of Radiology: Artificial Intelligence, Vice-chair of the SIIM ML Committee, and member of the RSNA AI Committee and RSNA Radiology Informatics Council; and serves on the Data Safety Monitoring Board for the LuANA Trial. **F. Nensa** serves as Associate Editor for Investigative Radiology, Section Editor (AI) for European Journal of Radiology, and Editor for European Journal of Radiology Artificial Intelligence. **J.N. Kather** provides consulting services for AstraZeneca and Bioptimus; holds shares in StratifAI, Synagen, and Spira Labs; has received institutional research grants from GSK and AstraZeneca; and has received honoraria from AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. J. N. Kather is also supported by the German Federal Ministry of Research, Technology and Space BMFTR (Come2Data, 16DKZ2044A; NextBIG, 01ZU2402A), the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/2 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden, the German Academic Exchange Service DAAD (SECAI, 57616814), and the European Research Council ERC (NADIR, 101114631). **L. Moy** serves on the ACR Data Safety Monitoring Board and the Society of Breast Imaging Board of Trustees; is on the editorial board of JMRI; receives a Siemens Research Grant; and receives personal fees from Bracco and Medscape. **M. Dietzel** serves as Editor-in-Chief of European Journal of Radiology Artificial Intelligence and Deputy Editor-in-Chief of European Journal of Radiology. **M. Huisman** has received speaker honoraria from Canon, Sonoskills, and AbbVie; serves on the Medical Advisory Board of xAID LLC; received a grant reviewing honorarium from the NN Foundation; received support for travel from ESR/EuSoMII; holds leadership roles including EuSoMII Vice President Elect (2025–26), member

of the ESR eHealth & Informatics Subcommittee, AI committee member for FMS and UEMS, Chair of the AI Task Force Biomedical Alliance, and Deputy Editor of Radiology: Artificial Intelligence. **S. Faghani** serves as Associate Editor of Radiology: Artificial Intelligence. **S. Tayebi Arasteh** serves as an editorial board member for Communications Medicine and European Radiology Experimental, and as a trainee editorial board member for Radiology: Artificial Intelligence. **W. Kim** serves as Chief Strategy Officer and CMIO at HOPPR; CMO at the American College of Radiology Data Science Institute; is on the Advisory Boards of Alara Imaging, Braid Health, ImageBiopsy Lab, and Luxsonic Technologies; is an Advisor and Shareholder at Rad AI; and is a Consultant for Hyperfine Research and Philips. **B. Kocak** served as Section Editor for Diagnostic and Interventional Radiology during the conduct of this study. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. All other authors declare no conflict of interest.

Author details

[1]Uskudar State Hospital, Department of Radiology, Istanbul, Türkiye

[2]Division of Diagnostic and Interventional Neuroradiology, Department of Radiology, University Hospital Basel, Basel, Switzerland

[3]University Children's Hospital Basel, Department of Pediatric Radiology, Basel, Switzerland

[4]Institute for Diagnostic and Interventional Radiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland

[5]Stanford University, Department of Radiology, Stanford, United States of America

[6]Technical University of Munich, School of Medicine and Health, Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, TUM University Hospital, Munich, Germany

[7]Technical University of Munich, School of Medicine and Health, Department of Cardiovascular Radiology and Nuclear Medicine, German Heart Center, TUM University Hospital, Munich, Germany

[8]Department of Medicine, Surgery, and Dentistry, University of Salerno, Baronissi, Italy

[9]Stanford University, Department of Biomedical Data Science, Stanford, United States of America

[10]University of California, San Francisco, Department of Radiology and Biomedical Imaging, San Francisco, California, United States of America

[11]King's College London, School of Biomedical Engineering and Imaging Sciences, LIHE, The London Institute for Healthcare Engineering, London, United Kingdom

[12]Department of Advanced Biomedical Sciences, University of Naples "Federico II", Naples, Italy

[13]Charité-Universitätsmedizin Berlin, Department of Neuroradiology, Humboldt-Universität zu Berlin, Freie Universität Berlin, Berlin Institute of Health, Berlin, Germany

[14]Yale University, Yale School of Medicine, Department of Radiology and Biomedical Imaging, New Haven, Connecticut, United States of America

[15]Mayo Clinic, Department of Radiology, Rochester, United States of America

[16]Lille University Hospital, Department of Neuroradiology, Lille, France

[17]University of Pennsylvania, Philadelphia, PA, United States of America

[18]University of Ulsan College of Medicine, Department of Radiology and Research Institute of Radiology, Asan Medical Center, Seoul, Republic of Korea

[19]University Medical Center Mainz, Department of Radiology, Mainz, Germany

[20]Royal Marsden Hospital, Department of Radiology and AI Imaging Hub, Sutton, United Kingdom

[21]Institute of Cancer Research, Division of Radiotherapy and Imaging, Sutton, United Kingdom

[22]Artificial Intelligence and Translational Imaging (ATI) Lab, Department of Radiology, University of Crete School of Medicine, Heraklion, Greece

[23]Medical Center - University of Freiburg Faculty of Medicine, Department of Diagnostic and Interventional Radiology, Freiburg, Germany

[24]St. Michael's Hospital, Department of Medical Imaging, Unity Health Toronto, Toronto, Canada

[25]University of Toronto Temerty Faculty of Medicine, Department of Medical Imaging, Toronto, Canada

[26]Universidade Federal de São Paulo, Department of Diagnostic Imaging, São Paulo, Brazil

[27]Institute for Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany

[28]Shanghai Key Laboratory of Magnetic Resonance, Institute of Magnetic Resonance and Molecular Imaging in Medicine, East China Normal University, China

[29]Informatics Institute, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland

[30]University of Geneva, Geneva, Switzerland

[31]The Sense Research and Innovation Center, Sion & Lausanne, Switzerland

[32]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine, TUD Dresden University of Technology, Dresden, Germany

[33]Department of Medicine I, Faculty of Medicine, TUD Dresden University of Technology, Dresden, Germany

[34]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

[35]Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, United Kingdom

[36]Department of Imaging, Tongren Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[37]Shanghai Key Laboratory of Flexible Medical Robotics, Tongren Hospital, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China

[38]Digital Surgery Lab - Breast Cancer Research Program, Champalimaud Foundation, Lisbon, Portugal

[39]University of Lisbon Faculty of Medicine, Department of Radiology, Lisbon, Portugal

[40]Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, Germany

[41]Champalimaud Foundation, Lisbon, Portugal

[42]Universitat de Barcelona, Departament de Matemàtiques i Informàtica, Barcelona, Spain

[43]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[44]Hellenic Mediterranean University, Department of Electrical and Computer Engineering, Heraklion, Crete, Greece

[45]Computational BioMedicine Laboratory, Institute of Computer Science, Foundation for Research and Technology (FORTH), Crete, Greece

[46]German Cancer Research Center (DKFZ), Division of Intelligent Medical Systems, Heidelberg, Germany

[47]National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and University Hospital Heidelberg, Heidelberg, Germany

[48]Heidelberg University Hospital, Surgical Clinic, Surgical AI Research Group, Heidelberg, Germany

[49]Heidelberg University, Faculty of Mathematics and Computer Sciences, Heidelberg, Germany

[50]Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates

[51]New York University Grossman School of Medicine, United States of America

[52]Medical Imaging Department and Biomedical Imaging Research Group at Hospital Universitario y Politécnico La Fe and Health Research Institute, Valencia, Spain

[53]Institute of Radiology, Universitätsklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

[54]Radboud University Medical Center, Department of Radiology and Nuclear Medicine, Nijmegen, The Netherlands

[55]LMU University Hospital, Department of Radiology, Munich, Germany

[56]Munich Center for Machine Learning (MCML), Munich, Germany

[57]relAI – Konrad Zuse School of Excellence in Reliable AI, Munich, Germany

[58]Computational Clinical Imaging Group, Champalimaud Research, Lisbon, Portugal

[59]AI Hub, Royal Marsden Hospital, Sutton, United Kingdom

[60]UT Southwestern Medical Center, Department of Radiology, Dallas, TX, United States of America

[61]Medical Anomaly Detection (MANO) Group, Computational Imaging Research (CIR), Department of Biomedical Imaging and Image-guided Therapy, Medical University of Vienna, Austria

[62]Comprehensive Center for AI in Medicine, Medical University of Vienna, Vienna, Austria

[63]University Hospital Regensburg, Department of Diagnostic and Interventional Radiology, Regensburg, Germany

[64]University of Pennsylvania, Department of Radiology, Philadelphia, United States of America

[65]Radiology Informatics Lab, Mayo Clinic, Department of Radiology, Rochester, United States of America

[66]Lab for AI in Medicine, Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany

[67]Stanford University School of Medicine, Department of Urology, Stanford, United States of America

[68]Rajiv Gandhi Cancer Institute and Research Center, Department of Radiology, New Delhi, India

[69]HOPPR, Illinois, United States of America

[70]American College of Radiology Data Science Institute, Virginia, United States of America

[71]Palo Alto VA Medical Center, California, United States of America

[72]Basaksehir Cam and Sakura City Hospital, Department of Radiology, Istanbul, Türkiye

References

1. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319-328. [Crossref]
2. Busch F, Hoffmann L, Rueger C, et al. Current applications and challenges in large language models for patient care: a systematic review. *Commun Med (Lond)*. 2025;5(1):26. [Crossref]
3. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform*. 2024;151:104620. [Crossref]
4. Bagde H, Dhopte A, Alam MK, Basri R. A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. *Heliyon*. 2023;9(12):e23050. [Crossref]
5. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. *BJOG*. 2024;131(3):378-380. [Crossref]
6. Wang L, Li J, Zhuang B, et al. Accuracy of large language models when answering clinical research questions: systematic review and network meta-analysis. *J Med Internet Res*. 2025;27:e64486. [Crossref]
7. Huo B, Boyle A, Marfo N, et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw Open*. 2025;8(2):e2457879. [Crossref]
8. Comeau DS, Bitterman DS, Celi LA. Preventing unrestricted and unmonitored AI experimentation in healthcare through transparency and accountability. *NPJ Digit Med*. 2025;8(1):42. [Crossref]
9. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res*. 2023;25:e50638. [Crossref]
10. Liu NF, Lin K, Hewitt J, et al. Lost in the middle: how language models use long contexts. *arXiv*. Preprint posted online November 20, 2023. [Crossref]
11. He J, Rungta M, Koleczek D, Sekhon A, Wang FX, Hasan S. Does prompt formatting have any impact on LLM performance? *arXiv*. Preprint posted online November 15, 2024. [Crossref]
12. Moëll B, Sand Aronsson F. Harm reduction strategies for thoughtful use of large language models in the medical domain: perspectives for patients and clinicians. *J Med Internet Res*. 2025;27:e75849. [Crossref]
13. Choudhury A, Chaudhry Z. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. *J Med Internet Res*. 2024;26:e56764. [Crossref]
14. Gu B, Desai RJ, Lin KJ, Yang J. Probabilistic medical predictions of large language models. *npj Digital Medicine*. 2024;7(1):367. [Crossref]
15. Akinci D'Antonoli T, Bluethgen C, Cuocolo R, Klontzas ME, Ponsiglione A, Kocak B. Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. *Diagn Interv Radiol*. 2025. [Crossref]
16. Kocak B. REFINE [Internet]. Open Science Framework; 2025 Aug 19. [Crossref]
17. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*. 2024;25(10):865-868. [Crossref]
18. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31(1):60-69. [Crossref]
19. Tripathi S, Alkhulaifat D, Doo FX, et al. Development, evaluation, and assessment of large language models (DEAL) checklist: a technical report. *NEJM AI*. 2025;2(6):AIp2401106. [Crossref]
20. Xie Q, Chen Q, Chen A, et al. Medical foundation large language models for comprehensive text analysis and beyond. *NPJ Digit Med*. 2025;8(1):141. [Crossref]
21. AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J Med Internet Res*. 2024;26:e59505. [Crossref]
22. Shen Y, Xu Y, Ma J, et al. Multi-modal large language models in radiology: principles, applications, and potential. *Abdom Radiol (NY)*. 2025;50(6):2745-2757. [Crossref]
23. Maaz S, Palaganas JC, Palaganas G, Bajwa M. A guide to prompt design: foundations and applications for healthcare simulationists. *Front Med (Lausanne)*. 2024;11:1504532. [Crossref]
24. Patil R, Heston TF, Bhuse V. Prompt engineering in healthcare. *Electronics*. 2024;13(15):2961. [Crossref]
25. Schader L, Song W, Kempker R, Benkeser D. Don't let your analysis go to seed: on the impact of random seed on machine learning-based causal inference. *Epidemiology*. 2024;35(6):764-778. [Crossref]
26. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nat Med*. 2025;31(3):943-950. [Crossref]
27. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018. [Crossref]
28. Ho CN, Tian T, Ayers AT, et al. Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. *BMC Med Inform Decis Mak*. 2024;24(1):357. [Crossref]
29. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell*. 2020;2(2):e200029. [Crossref]
30. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell*. 2024;6(4):e240300. [Crossref]
31. Martindale APL, Llewellyn CD, de Visser RO, et al. Concordance of randomised controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines. *Nat Commun*. 2024;15(1):1619. [Crossref]
32. 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*. 2024;385:e078378. [Crossref]
33. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. *Nat Med*. 2025;31(10):3283-3289. [Crossref]
34. The CHART Collaborative. Reporting guideline for chatbot health advice studies: the CHART statement. *JAMA Netw Open*. 2025;8(8):e2530220. [Crossref]
35. Johri S, Jeong J, Tran BA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nat Med*. 2025;31(1):77-86. [Crossref]
36. Cacciamani GE, Eppler MB, Ganjavi C, et al. Development of the ChatGPT, Generative Artificial Intelligence and Natural Large Language Models for Accountable Reporting and Use (CANGARU) Guidelines. *arXiv*. Preprint posted online July 18, 2023. [Crossref]
37. Park SH, Suh CH, Lee JH, et al. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM): 2025 updates. *Korean J Radiol*. 2025;26(12):1123-1132. [Crossref]
38. Qi X, Zeng Y, Xie T, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv*. Preprint posted online October 5, 2023. [Crossref]
39. Liu Y, He H, Han T, et al. Understanding LLMs: a comprehensive overview from training to inference. *arXiv*. Preprint posted online January 6, 2024. [Crossref]

40. Naveed H, Khan AU, Qiu S, et al. A Comprehensive overview of large language models. *arXiv*. Preprint posted online October 17, 2024. [Crossref]

41. Djuhera A, Kadhe SR, Zawad S, Ahmed F, Ludwig H, Boche H. Fixing it in post: a comparative study of LLM post-training data quality and model performance. *arXiv*. Preprint posted online October 27, 2025. [Crossref]

42. Fraser KC, Dawkins H, Nejadgholi I, Kiritchenko S. Fine-tuning lowers safety and disrupts evaluation consistency. *arXiv*. Preprint posted online June 20, 2025. [Crossref]

43. Tarabanis C, Zahid S, Mamalis M, Zhang K, Kalampokis E, Jankelson L. Performance of publicly available large language models on internal medicine board-style questions. *PLOS Digit Health*. 2024;3(9):e0000604. [Crossref]

44. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2024;30(2):80-90. [Crossref]

45. Li L, Sleem L, Gentile N, Nichil G, State R. Exploring the impact of temperature on large language models: hot or cold? *Procedia Comput Sci*. 2025;264:242-251. [Crossref]

46. Kaddour J, Harris J, Mozes M, Bradley H, Raileanu R, McHardy R. Challenges and applications of large language models. *arXiv*. Preprint posted online July 19, 2023. [Crossref]

47. Sahoo SS, Plasek JM, Xu H, et al. Large language models for biomedicine: foundations, opportunities, challenges, and best practices. *J Am Med Inform Assoc*. 2024;31(9):2114-2124. [Crossref]