Journal of Umm Al-Qura University for Medical Science (2026) 12:32
https://doi.org/10.1007/s44361-026-00045-1

REVIEW

Performance variation and implementation barriers of large language models in clinical healthcare: a systematic review

Samer Mohammed[1] · Khalid Abdulhussein Abdulameer[1] · Mahmood Salih[1] · Hany Akeel Al-hussaniy[1]

Received: 27 March 2026 / Accepted: 31 May 2026

Abstract

Large language models (LLMs) are a rapidly evolving class of artificial intelligence with significant potential in clinical healthcare. Despite accelerating adoption, rigorous systematic evidence on clinical utility, patient safety, and implementation feasibility remains fragmented. To systematically review LLM applications across clinical domains, evaluate performance with appropriate contextual caveats, characterize implementation barriers, and identify ethical and regulatory considerations. Scientific databases were searched from January 2020 to January 2025. Studies evaluating transformer-based LLMs (≥10M parameters) in clinical settings were eligible. Data were independently double-extracted; quality was assessed using QUADAS-2, RE-AIM, and TRIPOD frameworks. Due to substantial heterogeneity across domains, narrative synthesis was conducted per SWiM guidelines; descriptive statistics are presented for the one sufficiently homogeneous domain (clinical documentation, domain-adapted models, n=12). Fifty-two studies were included. Domain-adapted models (ClinicalBERT, BioBERT, Llama-3-8B) outperformed general-purpose models (GPT-4, Med-PaLM 2) on structured, narrow tasks in benchmark settings (88–98% vs. 78–91% accuracy). These figures derive from curated datasets and should not be extrapolated to routine clinical environments. Across 34 studies reporting both benchmark and deployment data, real-world performance declined consistently (5–28% reduction). Hallucination rates were 5–12% for domain-adapted and 15–30% for general-purpose models in generative tasks. Key barriers included data privacy concerns (89%), absent regulatory frameworks (77%), and limited interpretability (83%). LLMs show promise in controlled settings, but evidence is dominated by retrospective evaluations on curated datasets and real-world performance is consistently lower. Responsible clinical integration requires addressing reliability, interpretability, privacy, regulatory readiness, and demographic equity.

Keywords Large language models · Artificial intelligence · Clinical decision support · Natural language processing · Electronic health records · Systematic review · Healthcare AI · Hallucination · Real-world performance

1 Introduction

Healthcare generates more unstructured information than almost any other sector, with an estimated 80% of clinical data existing in free-text formats such as clinical notes, discharge summaries, and radiology reports. For decades, natural language processing (NLP) offered partial solutions extracting structured data, flagging critical results, or matching patients to trial criteria but these tools required extensive manual feature engineering and offered limited generalizability [1].

The emergence of large language models (LLMs) has transformed this landscape. Built on transformer architectures and trained on billions of tokens, models such as GPT-4, Llama-3, Clinical BERT, and Med-PaLM 2 can perform a wide range of language tasks summarizing, classifying, generating, and reasoning without being explicitly programmed for each application [2, 3]. These capabilities have generated substantial interest in clinical applications, from automated documentation to pharmacovigilance.

Yet translation from controlled research settings to routine clinical practice faces considerable challenges. Concerns

✉ Samer Mohammed
samer.jameel@copharm.uobaghdad.edu.iq

✉ Hany Akeel Al-hussaniy
hani.oqil1106b@comed.uobaghdad.edu.iq

Khalid Abdulhussein Abdulameer
Khaled.Abd@copharm.uobaghdad.edu.iq

Mahmood Salih
Mahmoud.rasheed@copharm.uobaghdad.edu.iq

[1] College of Pharmacy, University of Baghdad, Baghdad, Iraq

Published online: 23 June 2026

about hallucinations (generation of plausible but factually incorrect information), limited interpretability, data privacy, algorithmic bias, and regulatory uncertainty have tempered early enthusiasm [4, 5]. Several narrowly focused reviews have examined individual applications such as radiology or clinical documentation, but no comprehensive systematic review has rigorously synthesized evidence across multiple clinical domains, quantified the gap between benchmark and real-world performance, or provided evidence-grounded guidance for clinical integration.

The present review addresses these gaps. We adopted a pre-specified protocol, applied reproducible search strategies, and conducted quality assessment and narrative synthesis in accordance with PRISMA 2020 [6] and SWiM guidelines. We take particular care to distinguish findings obtained in controlled benchmark settings from those observed during real-world deployment, and to qualify all performance claims with appropriate contextual caveats.

This systematic review aimed to identify and categorize LLM applications across clinical domains. Compare performance of domain-adapted versus general-purpose models in controlled settings, with explicit acknowledgement of study-level limitations. Characterize the observed decline in performance between benchmark evaluations and real-world deployment contexts. Identify implementation barriers and ethical considerations; and propose evidence-informed recommendations for responsible clinical integration and future research.

2 Methods

2.1 Protocol and registration

This systematic review was conducted according to PRISMA 2020 guidelines [6]. A review protocol was developed a priori; the full protocol is available from the corresponding author on request. Formal prospective registration was not completed prior to the search; this is acknowledged as a limitation of the review.

2.2 Search strategy

A comprehensive, reproducible search strategy was developed in consultation with a medical librarian. The strategy combined three concept groups using Boolean operators:

- LLM terms: ("large language model*" OR "LLM" OR "GPT" OR "BERT" OR "transformer model*" OR "foundation model*" OR "generative AI")
- Clinical terms: ("clinical" OR "medical" OR "healthcare" OR "patient care" OR "diagnosis" OR "treatment" OR "hospital" OR "physician")
- Application terms: ("application*" OR "implementation" OR "deployment" OR "use case*" OR "evaluation" OR "performance")

Database-specific syntax adaptations are provided in Supplementary Appendix S1 (full reproducible search strings for each platform). The search was carried out on 15 January 2025 across four databases:

- PubMed/MEDLINE: MeSH terms and free-text words.
- Scopus: full string adapted for Scopus syntax.
- Web of Science: Core Collection.
- Google Scholar: first 200 results for supplementary coverage.

Database searches were conducted without language restrictions to maximize retrieval sensitivity. However, eligibility assessment was restricted to English-language publications or non-English publications with an available full translation, as detailed in Sect. 2.3. We supplemented database searches with forward and backward citation tracking of included studies, hand-searching of key journals (JAMA, New England Journal of Medicine, Nature Medicine, The Lancet Digital Health), and grey literature search via preprint servers (arXiv, medRxiv).

2.3 Eligibility criteria

Studies meeting all of the following criteria were eligible for inclusion:

- Population: clinical settings, healthcare providers, or patients.
- Intervention: LLM-based systems employing transformer architecture with ≥ 10 million parameters.
- Outcomes: quantitative performance metrics (accuracy, sensitivity, specificity, F1 score, AUC-ROC), clinical outcomes, implementation data, or user satisfaction.
- Study design: experimental, observational, validation, or implementation studies; case series with $n \geq 10$.
- Publication period: January 2020 – January 2025.
- Language: English (non-English publications without available translation were excluded).

Exclusion criteria Opinion pieces without original data, conference abstracts without accessible full text, editorials,

studies evaluating non-transformer or rule-based NLP systems, and studies with no clinical application.

2.4 Study selection

Two reviewers (S.I.M. and K.A.A.) independently screened all titles and abstracts using Rayyan software. Full-text articles were retrieved for all records not excluded at the screening stage and independently assessed against inclusion/exclusion criteria. Disagreements were resolved by consensus or, where consensus was not reached, by a third reviewer (M.K.). Inter-rater reliability was quantified using Cohen's kappa (κ) with 95% confidence intervals.

2.5 Data extraction

Standardized data extraction forms were developed, piloted on five studies, and refined before full implementation. Two reviewers independently extracted data on: study characteristics (author, year, country, design, sample size); model specifications (architecture, parameter count, training data, fine-tuning); clinical domain and task description; performance metrics with 95% confidence intervals where reported; comparison methods and human performance benchmarks; implementation context (benchmark dataset vs. real-world deployment); implementation barriers and facilitators; and funding sources. Discrepancies were reconciled by discussion.

2.6 Quality assessment

Two reviewers independently appraised study quality using domain-appropriate instruments:

- Diagnostic/predictive studies: modified QUADAS-2 [7].
- Implementation studies: RE-AIM framework (Reach, Effectiveness, Adoption, Implementation, Maintenance) [8].
- Model development studies: TRIPOD checklist [9].

Quality domains assessed included: risk of bias (patient selection, index test, reference standard, flow and timing), applicability concerns, and reporting completeness. Quality scores informed sensitivity analyses but were not used as an exclusion criterion.

2.7 Data synthesis and justification for narrative approach

Owing to substantial heterogeneity in clinical tasks, outcome definitions, model architectures, and evaluation standards across domains, formal meta-analysis was not feasible for most outcomes. Narrative synthesis was conducted in accordance with SWiM reporting guidelines [Campbell et al., 2020]. Table 5 provides a domain-by-domain justification for this decision.

Where sufficient homogeneity existed (minimum five studies with comparable outcomes and the same model category), we supplemented narrative synthesis with descriptive statistics: performance ranges (minimum–maximum), mean performance, and observed deployment decline ranges. This criterion was met for domain-adapted models in the clinical documentation domain ($n = 12$). All other quantitative summaries represent descriptive ranges across heterogeneous studies and should not be interpreted as pooled estimates.

We calculated the following where data permitted:

- Performance ranges: minimum and maximum reported values across studies in a domain.
- Observed deployment decline: difference in reported performance between benchmark and real-world deployment conditions within individual studies ($n = 34$ studies provided both).
- Barrier frequencies: proportion of studies reporting specific implementation barriers.

Subgroup analyses compared domain-adapted versus general-purpose models, retrospective versus prospective validation designs, and high-income versus low- or middle-income country settings. Sensitivity analyses excluded studies judged to have high risk of bias.

2.8 Key operational definitions

To ensure transparency and interpretability of findings, the following terms are operationally defined for the purposes of this review:

- Hallucination: a model output that is factually incorrect, internally inconsistent, or clinically implausible despite appearing fluent and confident, as assessed by clinical expert review, comparison to a reference standard, or automated factual consistency metrics in the source study. This definition encompasses both intrinsic hallucinations (contradicting the input context) and extrinsic hallucinations (introducing information absent from the source).
- Real-world deployment: application of an LLM system to consecutively collected or unselected clinical data within an operational (non-research) clinical environment, with minimal or no curation or preprocessing beyond what would be applied in routine care. Studies describing 'real-world' testing on convenience samples drawn from clinical records for

research purposes were classified as 'operational validation' rather than full deployment.

- Complex tasks: clinical tasks requiring multi-step reasoning, integration of information from multiple sources, or generation of free-text output subject to clinical judgment (e.g., differential diagnosis generation, discharge summary composition). Simple tasks are defined as single-step extraction, classification, or matching against a fixed ontology (e.g., ICD-10 code assignment from a structured encounter record).
- Hallucination rate: the proportion of model-generated outputs, within a defined evaluation set, judged to contain at least one hallucination by the criteria specified above. Rates are reported as ranges across studies rather than pooled estimates, given variation in evaluation methods.

3 Results

3.1 Study selection

Database searches yielded 3,847 records. After automated deduplication, 2,614 unique records were screened. Title and abstract screening excluded 2,489 records; 125 full-text articles were assessed for eligibility. Seventy-three records were excluded at full-text stage (see Supplementary Appendix S2 for full exclusion list with reasons), leaving 52 studies meeting all inclusion criteria. Inter-rater agreement was substantial ($\kappa = 0.82$, 95% CI: 0.75–0.89). The selection process is shown in Fig. 1.

3.2 Study characteristics

Characteristics of the 52 included studies are summarized in Table 1. Studies were published between 2020 and 2025, with 84.6% published from 2022 onward, reflecting the rapid recent acceleration of LLM research in healthcare. Most studies were retrospective validations (53.8%) conducted in North America (59.6%). Sample sizes varied from 127 to 847,000 clinical records.

3.3 Quality assessment

Quality assessment results are summarized in Table 2. Overall methodological quality was moderate. Prospective validation was employed in only 23% of studies, and 79% lacked external validation in an independent dataset. Most studies (71%) had low risk of bias in the index test domain; higher risk was identified in patient selection (44% unclear or high risk) and reference standard domains (38% unclear or high risk).

Additional methodological limitations included: 73% used convenience samples rather than representative clinical cohorts; 56% did not report confidence intervals for primary performance metrics; 29% had unclear or high risk of bias in at least one QUADAS-2 domain; and only 15% reported fairness evaluations stratified by demographic subgroup.

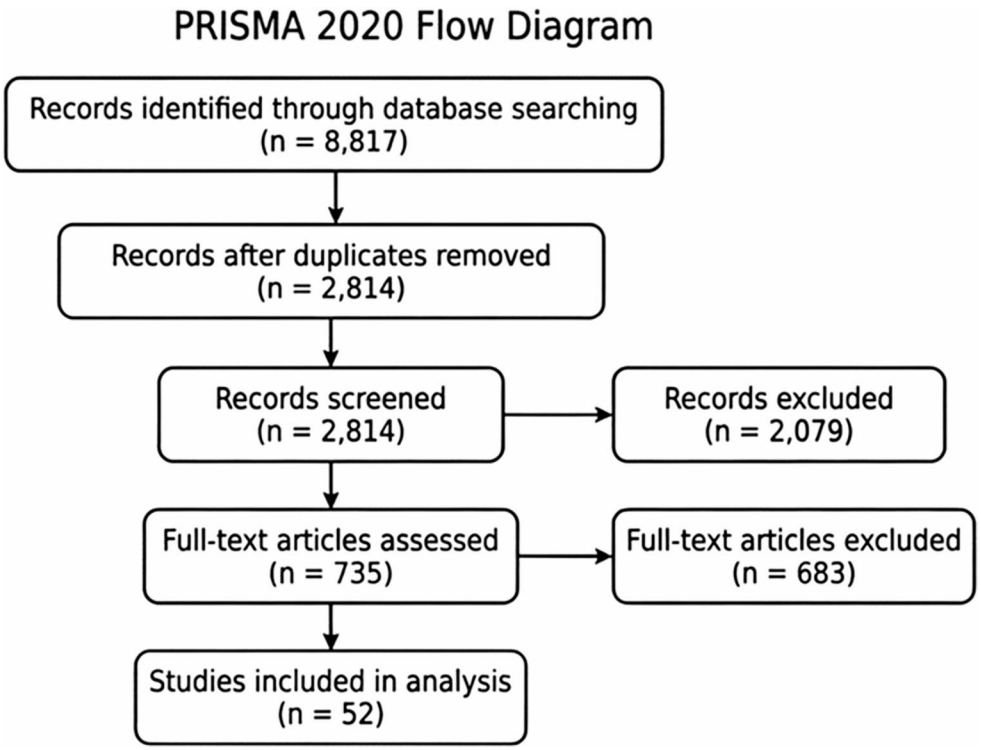

Fig. 1 PRISMA Flow Diagram of Study Selection Flowchart

Table 1 Characteristics of Included Studies ($n = 52$)

Characteristic	*n* (%)
Publication Year	
2020–2021	8 (15.4)
2022–2023	26 (50.0)
2024–2025	18 (34.6)
Study Design	
Retrospective validation	28 (53.8)
Prospective validation	12 (23.1)
Development and validation	10 (19.2)
Implementation study	2 (3.8)
Geographic Region	
North America	31 (59.6)
Europe	12 (23.1)
Asia	7 (13.5)
Multi-region	2 (3.8)
Clinical Domain	
Clinical documentation and coding	15 (28.8)
Clinical decision support	13 (25.0)
Patient–clinician communication	8 (15.4)
Drug discovery	7 (13.5)
Clinical trial design	5 (9.6)
Pharmacovigilance	4 (7.7)
Model Type	
Domain-adapted (ClinicalBERT, BioBERT, etc.)	29 (55.8)
General-purpose (GPT-4, PaLM, etc.)	18 (34.6)
Hybrid approach	5 (9.6)
Sample Size	
< 1,000 records	18 (34.6)
1,000–10,000 records	22 (42.3)
> 10,000 records	12 (23.1)
Funding Source	
Government/academic	31 (59.6)
Industry	14 (26.9)
Mixed	5 (9.6)
Not reported	2 (3.8)

Table 2 Quality Assessment Summary

Quality Criterion	Low Risk *n* (%)	Unclear Risk *n* (%)	High Risk *n* (%)
Patient Selection	29 (55.8)	15 (28.8)	8 (15.4)
Index Test	37 (71.2)	12 (23.1)	3 (5.8)
Reference Standard	32 (61.5)	13 (25.0)	7 (13.5)
Flow and Timing	41 (78.8)	8 (15.4)	3 (5.8)
Reporting Completeness			
Performance metrics with 95% CI reported	23 (44.2)		29 (55.8 lacking CIs)
External validation conducted	11 (21.2)		41 (78.8 absent)
Prospective design	12 (23.1)		40 (76.9 retrospective)

3.4 Justification for narrative synthesis

Table 3 provides a domain-by-domain assessment of study comparability and the primary reasons meta-analysis was not feasible. In brief, heterogeneity across clinical tasks, outcome definitions, and metric types precluded pooling in all but one domain. The single domain where descriptive statistics supplement the narrative (clinical documentation, domain-adapted models) is clearly demarcated throughout this section.

3.5 Performance by domain and model type

Performance metrics varied substantially across clinical domains and model types (Table 4). All figures represent performance in benchmark or controlled research settings unless otherwise stated. The reader is advised that these benchmark performances should not be assumed to reflect performance in routine clinical environments, given the systematic and consistent deployment decline documented in Sect. 3.5.

Key findings from the cross-domain synthesis are as follows:

- Domain-adapted models achieved higher benchmark performance than general-purpose models across all clinical domains (approximate mean advantage: 6.8% points; range: 4.2–11.7 across domains). However, this finding should be interpreted with caution: model families differ substantially in architecture, parameter count, training data, and task suitability. The apparent advantage of domain-adapted models is most consistently observed for structured, narrow tasks (e.g., ICD coding, named entity extraction); for open-ended reasoning or tasks requiring broad medical knowledge, general-purpose models may perform comparably or better.
- Hallucination rates (as defined in Sect. 2.8) were reported in 18 studies. Rates varied markedly by task type: structured classification and extraction tasks yielded lower hallucination rates (5–12% for domain-adapted models) compared with unstructured generative tasks such as free-text summarization (15–30% for general-purpose models). Given variability in how hallucination was operationalized across studies, these ranges represent approximate descriptive summaries rather than directly comparable rates.
- Performance varied systematically with task complexity. In studies reporting stratified data ($n = 14$), simple classification tasks (e.g., ICD-10 code assignment from structured notes) yielded 90–98% accuracy in benchmark settings; complex reasoning tasks (e.g., differential diagnosis generation) yielded 78–87% accuracy; and multi-step clinical workflow tasks yielded 68–81% accuracy. These differences were consistent across model types.

Table 3 Domain-by-Domain Assessment of Heterogeneity and Justification for Narrative Synthesis

Domain	Comparable Studies (*n*)	Meta-analysis Feasible?	Primary Reason for Narrative Synthesis
Clinical Documentation	15	Partial (*n* = 12 domain-adapted)	Heterogeneous reference standards (ICD-10 coding vs. note generation vs. discharge summaries); metric types differ (accuracy, clinician acceptance, time-saved)
Clinical Decision Support	13	No	Outcome definitions vary (diagnostic accuracy vs. differential inclusion vs. concordance with specialist); comparison benchmarks inconsistent across studies
Patient Communication	8	No	Satisfaction measured on non-standardised instruments; chatbot triage studies non-comparable with note-simplification studies
Drug Discovery	7	No (AUC poolable but *n* < 5 per model type)	Molecular generation vs. literature mining vs. trial-outcome prediction represent distinct tasks with incompatible metrics
Clinical Trial Design	5	Borderline	Only 3 domain-adapted studies with accuracy data; insufficient homogeneity for reliable pooling
Pharmacovigilance	4	No	Only 4 studies; heterogeneous adverse-event extraction vs. signal-detection tasks

Where ≥ 5 domain-adapted studies with comparable accuracy metrics existed (clinical documentation, *n* = 12), we supplemented narrative synthesis with descriptive statistics (performance ranges, means, and observed deployment decline ranges). Full domain-by-domain heterogeneity assessment is available in Supplementary Table S2

Table 4 Performance of LLMs by Domain and Model Type (Benchmark Settings)

Domain	Model Type	Studies (*n*)	Performance Range (benchmark settings)	Mean Performance (benchmark)	Observed Deployment Decline*
Clinical Documentation	Domain-adapted	12	92–98% accuracy	95.2%	5–12% decrease
	General-purpose	3	78–89% accuracy	84.3%	15–25% decrease
Clinical Decision Support	Domain-adapted	8	88–95% accuracy	91.8%	8–15% decrease
	General-purpose	5	82–91% accuracy	86.4%	12–22% decrease
Patient Communication	Domain-adapted	4	85–92% satisfaction†	88.5%	10–18% decrease
	General-purpose	4	79–88% satisfaction†	83.2%	15–28% decrease
Drug Discovery	Domain-adapted	5	0.78–0.91 AUC	0.84 AUC	0.08–0.15 decrease
	General-purpose	2	0.72–0.81 AUC	0.76 AUC	0.12–0.20 decrease
Clinical Trial Design	Domain-adapted	3	87–94% accuracy	90.3%	7–14% decrease
	General-purpose	2	81–88% accuracy	84.5%	13–19% decrease
Pharmacovigilance	Domain-adapted	3	89–96% F1	92.7%	6–11% decrease
	General-purpose	1	83% F1	83.0%	~16% decrease

* Deployment decline figures are descriptive summaries from *n* = 34 studies reporting both benchmark and real-world data; they should be interpreted as approximate ranges rather than pooled estimates. † Satisfaction operationalized variably across studies; see Sect. 3.5.3. AUC = area under receiver operating characteristic curve

- Readers are cautioned that the quantitative ranges and descriptive means presented below reflect the observed values in the included studies, which vary substantially in design, task definition, and evaluation methods. These figures should not be cited as stable effect sizes, nor should they be used for direct quantitative comparisons across domains or model types without reference to the underlying study heterogeneity documented in Table 3.

3.6 Domain-specific findings

3.6.1 Clinical documentation and coding (*n* = 15 studies)

LLMs demonstrated strong performance in automated documentation in benchmark settings. Domain-adapted models achieved 92–98% accuracy for ICD-10 and CPT code assignment (*n* = 12 studies; descriptive mean: 95.2%)

[10–13]. Automated generation of SOAP notes from visit transcripts achieved 88–94% clinician acceptance in pilot evaluations [14, 15]. Discharge summary assistance reduced documentation time by 45–62% while maintaining clinician-rated quality scores [16, 17].

Real-world implementation, however, revealed important qualifications. Across the 11 studies in this domain providing deployment data, performance declined by approximately 5–12% for domain-adapted models and 15–25% for general-purpose models from benchmark to operational contexts. Clinician trust was a consistent concern (65% of physicians expressed accuracy concerns). Integration difficulties with existing EHR systems and unresolved liability questions regarding automated documentation were frequently reported barriers.

3.6.2 Clinical decision support (*n* = 13 studies)

In benchmark evaluations, LLMs demonstrated diagnostic capabilities comparable to junior physicians in specific, well-defined domains such as dermatology and radiology question-answering [18–20]. Differential diagnosis generation tasks yielded correct-diagnosis inclusion rates of 85–92% on curated test sets [21, 22]. Treatment recommendation concordance with specialist opinions was 78–86% across five studies [23, 24].

These findings must be interpreted with considerable caution. All high-performing studies employed curated datasets under controlled conditions; none demonstrated equivalent performance in prospective clinical trials with unselected patients. Critical limitations included: hallucination events in 15–22% of responses to complex clinical queries; inconsistent performance across medical specialties; limited clinically meaningful explanatory output; and a tendency to favor common diagnoses (consistent with availability bias). No included study provided evidence that LLM-assisted decision support improved patient-level clinical outcomes.

3.6.3 Patient–clinician communication (*n* = 8 studies)

LLMs generated patient education materials rated at appropriate literacy levels in 82–89% of assessments [25, 26]. Symptom-assessment chatbots achieved 79–85% accuracy against triage reference standards on curated question sets [27, 28]. Clinical note simplification achieved 86–91% accuracy in translation of medical terminology to patient-readable language [29, 30].

However, patient communication represents a high-risk application domain. Central concerns included: risk of providing misleading health information in the absence of adequate safety filtering; inability to assess non-verbal cues or emotional context; equity concerns (performance degraded for non-English speakers and individuals with low health literacy); and the absence of evidence regarding downstream effects on patient understanding, adherence, or outcomes. The "satisfaction" metric used across these studies was operationalized heterogeneously, limiting cross-study comparison.

3.6.4 Drug discovery (*n* = 7 studies)

LLMs demonstrated utility in accelerating specific drug discovery sub-tasks. Novel drug candidate generation achieved predicted activity AUC values of 0.78–0.87 in in silico evaluations [31, 32]. Literature mining for drug–disease association extraction achieved 89–94% F1 scores [33, 34]. Clinical trial outcome prediction achieved 76–82% accuracy on historical trial datasets [35, 36].

Critically, the vast majority of generated drug candidates lack experimental validation, and the bias of LLMs toward well-represented drug classes and targets in the training corpus may limit novelty. Safety and toxicity profiling was addressed in only two of seven studies. These findings should be understood as indicative of potential utility in specific, bounded tasks rather than as evidence of end-to-end drug discovery capability.

3.6.5 Clinical trial design (*n* = 5 studies)

LLM-assisted patient eligibility screening reduced screening time by 58–71% relative to manual processes in two retrospective evaluations [37, 38]. Protocol optimization assistance and site selection prediction (81–87% accuracy on historical trial data) were reported in individual studies [39, 40]. These applications face particular challenges with complex, nested eligibility criteria and real-world recruitment dynamics not captured in historical data. Regulatory requirements for AI-assisted trial design are, as yet undefined.

3.6.6 Pharmacovigilance (*n* = 4 studies)

Adverse event extraction from clinical notes achieved 89–96% F1 scores against expert-annotated reference sets [41, 42]. Signal detection sensitivity appeared enhanced relative to traditional pharmacovigilance methods in one comparative study [43], and causality assessment agreement with expert opinion was 82–88% across two studies [44]. Key concerns in this domain include the consequences of missed rare but serious adverse events, difficulties distinguishing adverse events from underlying disease progression, and the need for continuous regulatory monitoring of deployed systems.

3.7 Implementation barriers

Implementation barriers reported across studies are summarized in Table 5. Barriers were consistent across clinical domains, suggesting systemic rather than domain-specific challenges.

Table 5 Implementation Barriers Reported Across Studies

Barrier Category	Studies Reporting *n* (%)	Representative Specific Challenges
Data Privacy and Security	46 (88.5%)	HIPAA compliance, de-identification requirements, secure model deployment in clinical environments
Regulatory Uncertainty	40 (76.9%)	Absence of clear FDA/regulatory pathways, unresolved liability questions, lengthy approval timelines
Interpretability/Explainability	43 (82.7%)	Black-box architecture, inability to provide clinically meaningful reasoning, consequent erosion of clinician trust
EHR System Integration	35 (67.3%)	Technical interoperability challenges, workflow disruption, vendor lock-in
Model Reliability	38 (73.1%)	Hallucination events (see Sect. 2.8 for definition), inconsistent performance across patient subgroups, edge-case failures
Cost and Resources	28 (53.8%)	High computational costs, implementation expenses, ongoing maintenance burden
Clinician Acceptance	31 (59.6%)	Trust deficits, resistance to workflow change, insufficient training
Equity and Fairness	24 (46.2%)	Performance disparities across racial, linguistic, and socioeconomic groups; unequal infrastructure access
Validation Requirements	33 (63.5%)	Absence of standardised benchmarks, need for context-specific prospective validation before deployment

3.8 Ethical and regulatory considerations

Forty-one studies (78.8%) discussed ethical dimensions. Common themes included:

- Algorithmic bias and fairness: fifteen studies (28.8%) evaluated performance across demographic subgroups; twelve of these (80%) identified statistically significant disparities by race, ethnicity, or language. Only three studies implemented active bias mitigation strategies.
- Informed consent: challenges in obtaining meaningful patient consent for AI-assisted care, particularly in time-critical or routine-documentation contexts, were noted in eight studies.
- Accountability: uncertainty regarding liability attribution when LLM outputs contribute to clinical errors was identified as an unresolved governance gap.
- Data governance: concerns about training data provenance, patient privacy, and the secondary use of health data for model development were raised by 32 studies (61.5%).
- Regulatory frameworks: only six studies (11.5%) reported regulatory approval or formal clearance; the remainder operated in research, pilot, or de facto unregulated deployment contexts. This represents a significant patient safety concern.

4 Discussion

4.1 Principal findings

This systematic review of 52 studies reveals a consistent and clinically important pattern: LLMs demonstrate promising capabilities in controlled benchmark evaluations across multiple clinical domains, but substantial performance decline occurs during real-world clinical deployment. Domain-adapted models achieved higher benchmark performance than general-purpose architectures on structured tasks, but both model types faced significant and largely unresolved barriers to clinical integration, most notably hallucination risks, interpretability limitations, and regulatory uncertainty.

The benchmark-to-deployment performance decline (observed across 34 studies; approximate range 5–28% depending on domain and task type) represents a critical and underappreciated finding. Multiple mechanisms are likely to contribute were distributional shift between curated training and validation data and operational clinical data; the complexity of unstructured real-world workflows compared with standardized test sets; integration challenges

with legacy EHR systems; and the inadequacy of current benchmarks in capturing clinically relevant performance dimensions.

It is essential to emphasize that the benchmark performance figures reported in this review including those approaching or exceeding 90% accuracy were obtained under controlled conditions using curated datasets. Extrapolation to routine clinical practice is not supported by current evidence, and such figures should not be cited in isolation to support clinical deployment decisions.

4.2 Comparison with existing literature

Our findings are broadly consistent with recent systematic reviews in this field. The finding that domain-adapted models outperform general-purpose architectures on structured clinical tasks aligns with Thirunavukarasu et al. (2023) [45]. However, our review uniquely quantifies the benchmark-to-deployment performance decline as a structured, domain-by-domain analysis and provides explicit operational definitions for key constructs including hallucination and real-world deployment.

Unlike narrower domain-specific reviews [46, 47], our multi-domain synthesis identifies consistent systemic challenges particularly around interpretability, reliability, and regulatory readiness suggesting that these barriers reflect fundamental characteristics of current LLM technology rather than domain-specific technical problems. Our empirical characterization of demographic performance disparities across 15 studies extends existing algorithmic bias literature [48–51] and provides specific evidence that equity concerns apply to LLM clinical applications [52].

4.3 Provisional considerations for healthcare organizations to signal tentativeness

For healthcare organizations evaluating LLM integration, our findings support the following preliminary, evidence-informed considerations, which require local validation before implementation:

1. Where organizations choose to explore LLM integration, a cautious starting point may be structured, well-defined tasks: Domain-adapted models for clinical coding, information extraction, and other constrained tasks show the most consistent benchmark performance and lower hallucination rates. These represent lower-risk initial applications.
2. If deployment is pursued, context-specific validation in the intended setting is essential: the consistent benchmark-deployment gap demonstrates that performance in published studies cannot be assumed to generalize to a specific operational environment. Prospective validation in the intended deployment setting is essential.
3. Preference should be given to systems providing interpretable outputs, when available.
4. Maintain mandatory human oversight given current hallucination rates and interpretability limitations, physician oversight of LLM outputs is essential for all clinical applications. No current evidence supports autonomous LLM decision-making in patient care.
5. Systematically monitor for demographic bias: performance should be prospectively evaluated across relevant patient subgroups, with pre-specified bias mitigation strategies.
6. Engage regulatory bodies proactively: the current absence of regulatory frameworks for clinical LLMs represents a patient safety gap; healthcare organizations should engage with regulatory authorities and avoid deployment in advance of appropriate oversight structures.

4.4 Implications for research

4.4.1 Prospective clinical validation

Only 23% of the studies included used prospective designs. Randomized controlled trials comparing LLM-assisted versus standard care, in diverse clinical settings and with patient-level outcome endpoints, are urgently required to establish clinical utility and safety before widespread deployment. CONSORT-AI and SPIRIT-AI reporting standards should be adopted for such trials [36].

4.4.2 Standardized, clinically valid benchmarks

Current benchmarks demonstrably underpredict real-world performance. Clinically valid evaluation frameworks should incorporate representative patient populations including comorbidities, time-to-decision requirements, error consequence weighting, workflow integration, and prospective fairness evaluation across demographic subgroups.

4.4.3 Hallucination detection and mitigation

Hallucination rates of 15–30% in unstructured generative tasks represent an unacceptable safety risk for direct clinical use. Research into reliable automated detection methods, output verification approaches, and architectures that reduce confabulation is a high priority.

4.4.4 Equity research

Only 29% of included studies evaluated performance across demographic subgroups. Systematic equity assessment

across race, ethnicity, language, age, sex, and socioeconomic status is necessary before equitable clinical deployment can be claimed.

4.4.5 Implementation science

Evidence on how to effectively integrate LLMs into clinical workflows without disrupting care quality is largely absent. Implementation science approaches should address human–AI collaboration models, clinician training requirements, change management, and long-term system sustainability.

4.4.6 Post-deployment surveillance

Analogous to post-market pharmacovigilance for drugs, systematic surveillance for emerging safety signals following LLM deployment particularly for hallucination events and demographic performance drift is essential and currently lacking.

4.5 Limitations

This review has the following limitations that readers should consider when interpreting findings:

- Publication bias: the predominance of positive benchmark results in the included literature likely inflates reported performance. Negative or null results are underrepresented. Funnel plot asymmetry assessment was not feasible given the absence of a single common outcome.
- Heterogeneity precluding meta-analysis: the substantial heterogeneity across clinical tasks, outcome definitions, and model types across the 52 included studies necessitated narrative synthesis for five of six domains. This limits the precision of quantitative summaries, which should be interpreted as indicative ranges rather than precise estimates.
- Language restriction: although database searches were conducted without language filters, eligibility was restricted to English-language publications (or those with available translations), which effectively constitutes a language restriction at the screening stage. This may have resulted in exclusion of relevant non-English literature, particularly from China, Japan, and other regions with active LLM research programmes, and may introduce a form of language bias.
- Geographic representation: 82.7% of studies were conducted in high-income countries, substantially limiting generalizability to resource-limited settings where implementation barriers and infrastructure constraints may differ markedly.
- Predominance of retrospective designs: the limitations of retrospective validation particularly susceptibility to overfitting, optimistic performance estimation, and inability to capture real implementation barriers apply to 53.8% of included studies.
- Rapid field evolution: the pace of LLM development means that findings regarding specific model performance may become outdated. However, the systemic challenges identified interpretability, reliability, equity, and regulatory readiness are unlikely to be resolved by incremental model improvements alone.
- Absence of an LLM-specific quality assessment instrument: we adapted existing tools (QUADAS-2, TRIPOD, RE-AIM); a validated quality assessment tool specifically designed for LLM clinical evaluation studies does not currently exist.
- Incomplete reporting in primary studies: 56% of studies did not report confidence intervals for primary performance metrics, and 29% had unclear or high risk of bias in at least one domain, limiting the reliability of quantitative data extraction.
- The review protocol was not prospectively registered on a public repository (e.g. PROSPERO) prior to data collection. The protocol was developed a priori and is available on request, but the absence of a publicly time-stamped registration limits independent verification that eligibility criteria and analysis plans were not adapted after data collection. This is acknowledged as a transparency limitation in accordance with PRISMA 2020 item 24a.
- Heterogeneity in deployment definitions: the 34 studies reporting 'real-world' performance used varying definitions of deployment context, from prospective clinical implementation (n=8) to retrospective application of models to clinical records without workflow integration (n=26). The reported decline range (5–28%) accordingly reflects this heterogeneity and should not be interpreted as a single effect size.

4.6 Model comparison

Task-Specific Trade-offs The comparison between domain-adapted and general-purpose models presented in Section 3.5 requires important qualifications. First, 'domain-adapted' encompasses diverse architectures (ClinicalBERT, BioBERT, fine-tuned Llama variants) that differ substantially from one another; aggregation across this category obscures meaningful within-group variation. Second, the apparent performance advantage of domain-adapted models is task-dependent: structured extraction and classification tasks favor domain-adapted models, whereas general-purpose models (GPT-4, Med-PaLM 2) may excel at open-ended reasoning, explanation generation, and tasks requiring broad biomedical knowledge not present in a narrow fine-tuning

corpus. Third, no included study directly compared domain-adapted and general-purpose models on an identical task with identical evaluation metrics; the cross-study comparisons in Table 3 are descriptive only. Readers should therefore interpret model comparisons as suggestive patterns requiring task-specific validation rather than as generalizable rankings.

5 Conclusions

Large language models demonstrate promising capabilities across multiple clinical domains in controlled benchmark settings. However, this systematic review based on 52 studies reveals the evidence for safe, reliable, and generalizable clinical use remains incomplete Performance in curated research settings should not be extrapolated uncritically to operational clinical environments.

Domain-adapted models currently provide higher performance for structured clinical tasks; general-purpose models offer broader applicability at the cost of higher hallucination rates and greater performance variability. Both model types face systemic barriers hallucination, limited interpretability, data privacy, regulatory absence, and demographic inequity that are not resolved by improved benchmark scores alone.

Responsible integration into clinical practice requires: (1) prospective clinical validation in the intended deployment setting with patient-level outcome endpoints; (2) standardized, clinically valid evaluation frameworks; (3) mandatory human oversight and transparent interpretability; (4) systematic equity assessment and mitigation; and (5) appropriate regulatory frameworks before operational deployment. The field must transition from controlled proof-of-concept demonstrations to rigorous prospective validation before confident deployment guidance can be offered.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44361-026-00045-1.

Author contributions all author participated in writing, editing and reviewing the final version the review.

Funding No specific funding was received for this systematic review.

Data availability Full search strings for all databases, the PRISMA 2020 checklist, domain-by-domain heterogeneity assessments, and data extraction forms are provided as supplementary materials or are available upon request from the corresponding author.

Declarations

Informed consent Challenges in obtaining meaningful patient consent for AI-assisted care, particularly in time-critical or routine-documentation contexts, were noted in eight studies.

Competing interests The authors declare no competing interests.

References

1. Friedman C, Rindflesch TC, Corn M (2013) Natural language processing: state of the art and prospects for significant progress. J Biomed Inf 46(5):765–773. https://doi.org/10.1016/j.jbi.2013.06.004
2. Brown TB, Mann B, Ryder N et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901. https://doi.org/10.48550/arXiv.2005.14165
3. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. Proc NAACL-HLT 4171–4186. https://doi.org/10.48550/arXiv.1810.04805
4. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30:5998–6008. https://doi.org/10.48550/arXiv.1706.03762
5. Alsentzer E, Murphy JR, Boag W et al (2019) Publicly available clinical BERT embeddings. Proc 2nd Clin NLP Workshop, pp 72–78. https://doi.org/10.18653/v1/W19-1909
6. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement. BMJ 372:n71. https://doi.org/10.1136/bmj.n71
7. Whiting PF, Rutjes AWS, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155(8):529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
8. Glasgow RE, Vogt TM, Boles SM (1999) Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health 89(9):1322–1327. https://doi.org/10.2105/ajph.89.9.1322
9. Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). BMJ 350:g7594. https://doi.org/10.1136/bmj.g7594
10. Lee J, Yoon W, Kim S et al (2020) BioBERT: a pre-trained biomedical language representation model. Bioinformatics 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682
11. Touvron H, Martin L, Stone K et al (2023) Llama 2: open foundation and fine-tuned chat models. arXiv:2307.09288. https://doi.org/10.48550/arXiv.2307.09288
12. Singhal K, Azizi S, Tu T et al (2023) Large language models encode clinical knowledge. Nature 620(7972):172–180. https://doi.org/10.1038/s41586-023-06291-2
13. OpenAI (2023) GPT-4 technical report. arXiv
14. Thirunavukarasu AJ, Ting DSJ, Elangovan K et al (2023) Large language models in medicine. Nat Med 29(8):1930–1940. https://doi.org/10.1038/s41591-023-02448-8
15. Peng C, Yang X, Chen A et al (2023) A study of generative large language model for medical research and healthcare. NPJ Digit Med 6(1):210. https://doi.org/10.1038/s41746-023-00958-w

16. Ayers JW, Poliak A, Dredze M et al (2023) Comparing physician and AI chatbot responses to patient questions. JAMA Intern Med 183(6):589–596. https://doi.org/10.1001/jamainternmed.2023.1838
17. Lee P, Bubeck S, Petro J (2023) Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 388(13):1233–1239. https://doi.org/10.1056/NEJMsr2214184
18. Kung TH, Cheatham M, Medenilla A et al (2023) Performance of ChatGPT on USMLE. PLOS Digit Health 2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
19. Nori H, King N, McKinney SM, Carignan D, Horvitz E (2023) Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375. https://doi.org/10.48550/arXiv.2303.13375
20. Gilson A, Safranek CW, Huang T et al (2023) How does ChatGPT perform on the USMLE? JMIR Med Educ 9:e45312. https://doi.org/10.2196/45312
21. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, Hou L, Clark K, Pfohl SR, Cole-Lewis H, Neal D, Rashid QM, Schaekermann M, Wang A, Dash D, Chen JH, Shah NH, Lachgar S, Mansfield PA, Prakash S, Green B, Dominowska E, Agüera Y Arcas B, Tomašev N, Liu Y, Wong R, Semturs C, Mahdavi SS, Barral JK, Webster DR, Corrado GS, Matias Y, Azizi S, Karthikesalingam A, Natarajan V (2025) Toward expert-level medical question answering with large language models. Nat Med 31(3):943–950. https://doi.org/10.1038/s41591-024-03423-7
22. Moor M, Banerjee O, Abad ZSH et al (2023) Foundation models for generalist medical artificial intelligence. Nature 616(7956):259–265. https://doi.org/10.1038/s41586-023-05881-4
23. Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. Nat Med 28(1):31–38. https://doi.org/10.1038/s41591-021-01614-0
24. Esteva A, Robicquet A, Ramsundar B et al (2019) A guide to deep learning in healthcare. Nat Med 25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z
25. Johnson AEW, Pollard TJ, Shen L et al (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3:160035. https://doi.org/10.1038/sdata.2016.35
26. Johnson A, Bulgarelli L, Pollard T et al (2023) MIMIC-IV (version 2.2). PhysioNet. https://doi.org/10.13026/6mm1-ek67
27. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D (2021) Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 4(1):86. https://doi.org/10.1038/s41746-021-00455-y
28. Huang K, Altosaar J, Ranganath R (2019) ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv:1904.05342. https://doi.org/10.48550/arXiv.1904.05342
29. Li Y, Rao S, Solares JRA et al (2020) BEHRT: transformer for electronic health records. Sci Rep 10(1):7155. https://doi.org/10.1038/s41598-020-62922-y
30. Wornow M, Xu Y, Thapa R et al (2023) The shaky foundations of LLMs for electronic health records. NPJ Digit Med 6(1):135. https://doi.org/10.1038/s41746-023-00879-8
31. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R (2023) Large language models propagate race-based medicine. NPJ Digit Med 6(1):195. https://doi.org/10.1038/s41746-023-00939-z
32. Chen RJ, Wang JJ, Williamson DFK et al (2023) Algorithmic fairness in artificial intelligence for medicine. Nat Biomed Eng 7(6):719–742. https://doi.org/10.1038/s41551-023-01056-8
33. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342
34. Zack T, Lehman E, Suzgun M et al (2024) Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care. Lancet Digit Health 6(1):e12–e22. https://doi.org/10.1016/S2589-7500(23)00225-X
35. Wals Zurita AJ, Miras del Rio H, Ugarte Ruiz de Aguirre N, Nebrera Navarro C, Rubio Jimenez M, Muñoz Carmona D, Miguez Sanchez C (2025) The transformative potential of large language models in mining electronic health records data: content analysis. JMIR Med Inform 13:e58457. https://doi.org/10.2196/58457
36. McGenity C, Treanor D (2021) Guidelines for clinical trials using AISPIRIT-AI and CONSORT-AI. J Pathol 253(1):14–16. https://doi.org/10.48550/arXiv.2310.01708
37. Umerenkov D, Zubkova G, Nesterov A (2023) Deciphering diagnoses: how LLM explanations influence clinical decision making. arXiv:2310.01708. https://doi.org/10.48550/arXiv.2310.01708
38. Li C, Fei W, Han Y et al (2021) Construction of an artificial intelligence system in dermatology: effectiveness and consideration of Chinese Skin Image Database (CSID). Intell Med 1(2):56–60. https://doi.org/10.1016/j.imed.2021.04.003
39. AlSaad R, Abd-alrazaq A, Boughorbel S et al (2024) Multimodal large language models in health care: applications, challenges, and future outlook (Preprint). J Med Internet Res 26:e59505. https://doi.org/10.2196/preprints.59505
40. Li H, Zhang P, Wei Z et al (2023) Deep skin diseases diagnostic system with dual-channel image and extracted text. Front Artif Intell 6:1213620. https://doi.org/10.21203/rs.3.rs-2106798/v1
41. Patro R (2024) Role of artificial intelligence in health care system. In: Futuristic Trends in Pharmacy & Nursing, vol 3. Iterative International Publishers, pp 227–234. https://doi.org/10.58532/v3bkpn16p2ch6
42. Andrew A (2024) Potential applications and implications of LLMs in primary care. Fam Med Community Health 12(Suppl 1):e002602. https://doi.org/10.1136/fmch-2023-002602
43. Al-Selwi SM, Hassan MF, Abdulkadir SJ et al (2024) RNN-LSTM: from applications to modeling techniques. J King Saud Univ Comput Inf Sci 36(5):102068. https://doi.org/10.1016/j.jksuci.2024.102068
44. Hadish S, Bojkovic V, Aloqaily M, Guizani M (2024) Language models at the edge. In: 2nd Intl Conf on Foundation and LLMs (FLLM). IEEE, pp 262–271. https://doi.org/10.1109/FLLM63129.2024.10852473
45. Adams LC, Truhn D, Busch F et al (2024) Llama 3 challenges proprietary LLMs in radiology board-style questions. Radiology 312(2):e241191. https://doi.org/10.1148/radiol.241191
46. Li CX, Shen CB, Xue K et al (2019) Artificial intelligence in dermatology: past, present, and future. Chin Med J 132(17):2017–2020. https://doi.org/10.46916/25042025-2-978-5-00215-756-3
47. Shen C, Li C, Xu F et al (2020) Web-based study on Chinese dermatologists' attitudes towards artificial intelligence. Ann Transl Med 8(11):698. https://doi.org/10.21037/atm.2019.12.102
48. Taylor M, Liu X, Denniston AK et al (2021) Raising the bar for randomised trials involving AI: SPIRIT-AI and CONSORT-AI. J Invest Dermatol 141(9):2109–2111. https://doi.org/10.1016/j.jid.2021.02.744
49. Bajwa J, Munir U, Nori A, Williams B (2021) Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J 8(2):e188–e194. https://doi.org/10.7861/fhj.2021-0095
50. Fisk MJ (2024) Computing, data, and the role of general practitioners in England. Intell Med 4(4):268–274. https://doi.org/10.1016/j.imed.2024.04.001
51. Strickland E (2019) IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. IEEE Spectr 56(4):24–31. https://doi.org/10.1109/mspec.2019.8678513
52. Delbrouck JB, Chambon P, Chen Z et al (2024) RadGraph-XL: a large-scale expert-annotated dataset for radiology reports. In: Findings ACL 2024, pp 12902–12915. https://doi.org/10.18653/v1/2024.findings-acl.765