Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085321

REVIEW

Large Language Models in Biomedical Text Summarization: A Systematic Review of Architectures, Evaluation Adequacy, and Clinical Readiness

Adel Assiri*

Department of Business Informatics, College of Business, King Khaled University, Abha, Saudi Arabia

*Corresponding Author: Adel Assiri. Email: adaseri@kku.edu.sa

Received: 09 May 2026; Accepted: 20 July 2026; Published: 15 September 2026

ABSTRACT: The emergence of Large Language Models (LLMs) has transformed biomedical text summarization, shifting research beyond conventional extractive approaches toward increasingly generative and agentic reasoning paradigms. However, rapid advances in model capabilities have exceeded current understanding of their methodological rigor, evaluation adequacy, and clinical readiness. This systematic review conducts a structured exploratory audit of Transformer- and LLM-based biomedical summarization systems to characterize reporting quality, validation practices, and translational maturity. Following PRISMA 2020 guidelines and the Population–Concept–Context framework, we systematically reviewed 178 original English-language studies published between January 2017 and March 2026. A multidimensional evaluation pipeline was developed, incorporating the Methodological Quality Score (MQS) to assess reporting transparency, the Evaluation Adequacy Score (EAS) to quantify validation depth, and the Clinical Readiness Level (CRL) to classify deployment maturity. The field demonstrated substantial recent growth, with 75.8% of included studies published since 2024. Decoder-only architectures represented the most frequently reported model family (50.6%), followed by hybrid or agentic approaches (26.4%). Although reported ROUGE and BERTScore values indicated strong technical performance, substantial heterogeneity across datasets, clinical domains, and summarization tasks limited direct comparison between architecture families. The reviewed studies showed high reporting transparency (median MQS = 7) but limited evaluation depth (median EAS = 0.46). Furthermore, 75.3% of studies remained at laboratory or technical validation stages (CRL 1–2), while only 24.7% achieved institutional feasibility (CRL ≥ 3). Lexical similarity metrics demonstrated limited alignment with clinical readiness, highlighting the need for factuality-centered evaluation and safety-oriented validation strategies. The integrated MQS–EAS–CRL framework provides a structured approach for assessing methodological maturity and translational readiness in biomedical summarization research. Future studies should prioritize robust factuality evaluation, verification mechanisms, and broader clinical validation across diverse healthcare settings.

KEYWORDS: Large language models; biomedical text summarization; evaluation adequacy; clinical readiness; hallucination taxonomy; ROUGE gap; agentic workflows; systematic review

1 Introduction

Biomedical and clinical text summarization, referring to the automatic creation of a short and accurate summary of a text such as an electronic health record, radiology report, discharge summary, or scientific literature, has become a critical need in healthcare informatics. Growth in EHR (Electronic Health Record) documentation and biomedical literature increases the cognitive burden on clinicians and risks delaying timely access to evidence [1–3]. Transformer- and Large Language Model (LLM)-based summarization has become an important technological tool for managing information overload [2,4].

Biomedical text summarization has evolved through various architectural generations. This evolution reflects a progressive transition from lexical compression techniques toward semantically informed generative systems capable of synthesizing clinically meaningful information from increasingly complex biomedical documents. Early extractive methods used statistical weighting schemes and lexical graph-based methods, such as Term Frequency–Inverse Document Frequency (TF–IDF), TextRank, and LexRank, to select salient sentences from source documents without creating new content or synthesizing underlying meaning [5–7]. Hybrid architectures then sought to integrate these statistical approaches with biomedical knowledge bases, using concept-mapping techniques such as MetaMap, to improve clinical concept representation. This was followed by a transition to encoder–decoder sequence-to-sequence (Seq2Seq) architectures using recurrent models, especially Long Short-Term Memory (LSTMs) and later Gated Recurrent Unit (GRUs). These enabled early practical neural abstractive summarization, but remained constrained by sequential processing and limited parallelism. Although vanishing gradients were a major structural issue for vanilla Recurrent Neural Network (RNNs), LSTMs and GRUs were designed to mitigate this through gating mechanisms; however, they still struggled with very long-range dependencies and generally lagged behind later Transformer architectures on this dimension [8–11]. These limitations motivated the emergence of domain-specific Transformer architectures, representing a major methodological shift in biomedical text summarization [12–14], and were followed by dedicated encoder-decoder models such as BART, T5, BioBART, and PEGASUS that greatly improved abstractive fluency on reported benchmarks. The current generation is dominated by powerful decoder-only LLMs, such as general domain models (Generative Pre-trained Transformer (GPT), Llama, and Mistral) and medical domain-specific models (BioGPT, BioMistral, and Med-PaLM) [15–17]. Recent hybrid and agentic designs, including retrieval-augmented generation (RAG) and multi-agent pipelines, represent the latest architectural evolution [18–20].

Transformer- and LLM-based models offer several advantages for biomedical summarization due to three core architectural attributes. First, multi-head self-attention enables contextual representation learning, as the model can capture bidirectional dependencies among tokens across the entire clinical document rather than only among co-occurring diagnostic codes, clinical abbreviations, and anatomical descriptions [21]. Biomedical Pre-trained Language Model (PLM)s and LLMs based on Transformer architectures, such as BioBERT, ClinicalBERT, and GPT-family, have been effective for biomedical summarization and related tasks due to their ability to capture long-distance relationships and rich semantics in biomedical text [22–24]. Second, long-context Transformer models based on sparse attention, such as Longformer, BigBird, Clinical-Longformer, and Clinical-BigBird, reduce the memory burden of full self-attention and extend the processable input length, which improves modeling of long clinical texts and other tasks requiring long-range dependencies [25,26]. Third, large autoregressive generative models can write and rewrite complex medical narratives into concise, coherent summaries and patient explanations, moving beyond extractive sentence selection [7–10]. These represent the trend of transitioning from extractive methods, recurrent Seq2Seq models, and the previous generation of domain-specific Transformers to instruction-tuned LLMs for summarization in the biomedical and clinical domains [11–15].

In recent studies, adapted LLM models have performed well on various benchmark tasks, including generating radiology impressions, writing discharge summaries, and synthesizing clinical progress notes. Despite promising computational outcomes, significant concerns remain about their reliability, robustness, and safety in real-world healthcare settings. For this reason, the clinical usability of such systems is still under investigation and discussion [27–29]. The most systematic review of LLM-based clinical text summarization highlighted methodological and corpus limitations, as the selected studies were all retrospective, focused on radiology and Intensive Care Unit (ICU) notes, and used Medical Information Mart for Intensive Care (MIMIC)-derived datasets [4,30,31]. More comprehensive evaluations of LLM applications in healthcare

domains show that decoder-only general-domain models dominate, while domain-specific and alternative architectures remain underrepresented [32–34]. A few systematic reviews have focused on providing a longitudinal architectural taxonomy or a quantitative cross-study comparison of model performance across biomedical text types [31,35,36].

Evaluation practices are seen as the main constraint of the current review studies. Most studies are internally validated using lexical overlap scores such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Bilingual Evaluation Understudy (BLEU), sometimes supplemented with BERTScore [37,38]. Expert reviewers consistently reported that biomedical summarization tools are still suffering from hallucinations, factual inconsistencies, and clinically misleading errors [37,39,40]. Failure analysis was identified in 20% of studies, and only 3% of clinical summarization studies were formally assessed for patient safety, with no type of evaluation bias identified in external validation studies (7%) [4]. Similar assessment of 761 LLM appraisals in clinical medicine supports that consideration of fairness, quantification of uncertainty, and deployment concerns are rarely evaluated [32,41]. Although this bottleneck is widely recognized, few systematic reviews have measured this limitation across studies with a consistent, repeatable scoring system. Generalizability is further constrained by current Dataset concentration, with most summarization studies on clinical datasets focused on English-language institutional datasets like MIMIC-III/IV, and few studies covering other specialties, care settings, languages, or low-resource environments [34,42]. In addition, independent replication is limited because code, prompts, and data-processing pipelines are largely unavailable, and further translational and governance barriers exacerbate these evaluation deficits [4,13].

Furthermore, there are several issues highlighted by recent reviews of existing LLMs in healthcare, including biased training data, privacy and security issues, opaque cloud-based implementations, and the production of highly plausible but incorrect outputs [33,38,43]. While LLMs can outperform human experts on specific benchmark tasks, qualitative evaluations still reveal hallucinations and minor factual inaccuracies that could negatively impact patient care [41]. In addition, recent studies have shown that the performance of smaller open-weight models can be similar to that of much larger proprietary models on specific clinical tasks, meaning that model size is not a good indicator of clinical utility [21,43]. These constraints also indicate that benchmark performance provides insufficient evidence of clinical effectiveness and highlight the challenges that remain in clinical readiness for systems such as summarization, decision support, and patient communication [44–47]. These limitations were summarized in this study with six specific gaps below, which motivated the present review:

1. Because of the broad scope of summarization techniques and the diversity of their clinical uses, existing reviews have not been able to cover all areas of medical summarization. This limits the usefulness of the results in addressing different healthcare problems and applications [4,40,48].
2. Lack of a formal architectural taxonomy: existing reviews focus primarily on models in narrative terms rather than categorizing architectural paradigms and adaptation strategies, which limits cross-study comparison [49–51].
3. Inconsistency in evaluation practices: Critical omissions in factuality, safety, bias assessment, and external clinical validation are found in existing reviews [46,52].
4. Limited technical development and clinical implementation, constrained by deployment-readiness barriers, operational limitations, governance requirements, and regulatory compliance [53,54].
5. Lack of empirical evidence regarding the interdependence of evaluation rigor and clinical readiness: There is limited research that has examined the exploratory association between the rigor of evaluation and clinical readiness, which represents a mutual reinforcement of quality standards that remains unexplored [47,53].

6. Existing reviews primarily offer descriptive overviews, highlighting the need for a structured audit infrastructure to characterize methodological advances [55,56]. Table 1 below provides a comparative analysis of selected recent existing studies.

Table 1: Comparative analysis of recent systematic reviews relative to the current study scope.

Study	Timeframe & Volume (N)	Scope, Focus & Methodology	Analyzed Dimensions & Synthesis Level	Limitations & Gaps	RO1 Taxonomy Alignment?	Standardized Quality Tools (MQS/EAS)	Clinical Maturity Staging (CRL)
[41]	2022–2024 N = 519	LLM testing in clinical knowledge, triage, and administration: A systematic review (PubMed/WoS).	Data types, task classes, NLP types, eval dimensions, and specialties. Descriptive Quantitative Mapping.	95% skip real patient data; heavy concentration, omits toxicity/robustness indices.	No. Static cross-sectional snapshot; lacks multi-era tracking.	No (predefined study component categorization only; no formal quality scale).	No (deployment noted only as an infrequently measured evaluation metric).
[45]	2017–mid, 2024 N = 132	LLM functionalities in clinical language understanding (NER, NLI, VQA). Comprehensive Literature Survey.	PLM-to-LLM trajectories, medical task benchmarking (MedQA, MedNLI), and thematic ethical mapping.	Primarily technical metric focus; lacks standardized methodological quality audits of the included literature	No. (Uses chronological size-progression rather than the 3-era longitudinal taxonomy).	No. (Discusses standard NLP metrics—ROUGE, BLEU, F1—but offers no formal quality scale).	No. (Lacks a formal implementation/readiness staging pipeline).
[57]	2023–2025 N = 168	Autonomous and deliberative medical agents. System overview and active pipeline screening.	Memory modules, tool APIs, and glass-box configurations. Structural System Synthesis.	Relies heavily on de-identified or synthetic text and has severe real-time validation gaps for safety edge cases.	Partial. Profiles agentic ecosystems but omits longitudinal multi-era tracing.	No (highlights structural evaluation gaps without offering a standardized tool).	No (Identifies workflow integrations theoretically, without a staging pipeline).
[2]	2022–2024 N = 48	Generative AI in medical education and training: A Non-PRISMA Narrative Review.	Examination datasets, score benchmarks, and general use paths. Non-Systematic Narrative Tracking.	High risk of selection bias; lacks reproducible search loops or technical verification benchmarks.	No. Confined to academic performance; lacks multi-era tracking.	No (relies on narrative trends instead of reproducible validation tools).	Policy Imprecision: Offers abstract ethical guidelines instead of verifiable safety protocols.
[38]	2017–2024 N = 40	LLMs in psychiatric screening and chatbots. PRISMA systematic review.	Psychiatric task pillars, model sizes, and clinical risk factors. Narrative Descriptive Tracking.	Acute deficit of expert-annotated or multilingual clinical datasets and high hallucination vulnerability.	No. It is segmented strictly by psychiatric tasks and lacks an overarching taxonomy.	No (narrative review; highlights the critical lack of verified training benchmarks).	No (groups baseline accuracy and risks without a formal staging pipeline).
[21]	1-Year Window Landscape	Short-term tracking of domain-specific medical LLMs. Narrative landscape overview.	Release cycles, clinical operations, and literature summarization. Descriptive Chronological Timeline.	Rapid model obsolescence; highly dependent on unverified vendor-reported metrics; and unverified safety.	Partial. Captures a brief 12-month snapshot but lacks long-term evolutionary metrics.	No (chronological tracking entirely reliant on raw vendor performance claims).	No (fails to verify real-world utility or map a standardized maturity pipeline).

(Continued)

Table 1 (continued)

Study	Timeframe & Volume (N)	Scope, Focus & Methodology	Analyzed Dimensions & Synthesis Level	Limitations & Gaps	RO1 Taxonomy Alignment?	Standardized Quality Tools (MQS/EAS)	Clinical Maturity Staging (CRL)
[4]	2023–2025 N = 64	State of Art (SOTA) validation in clinical-text summarization. PRISMA-Scoping Review (PRISMA-ScR).	Text compaction layouts, data criteria, and lexical parameters. Dataset Dependency Analysis.	Critically over-reliant on MIMIC-IV data; lacks out-of-distribution validation across diverse demographic groups.	Partial. Tracks summarization patterns retrospectively and misses cross-era structural tracing.	Audit Deficit: Identifies bias but provides no engineering protocol to measure or mitigate it.	No (methodological mapping is completely isolated from progressive maturity pipelines).
[58]	2020–2023 N = 58	Metric evaluation paradigms for medical-language models. Systematic Scoping Review.	Automated linguistic metrics, expert rubrics, and testing tracks. Descriptive Stagnation Framework.	It fails to resolve metric fragmentation and exposes rampant "Evaluation Anarchy" across research teams.	No. Confined strictly to metric tracking, it does not build a multi-era performance synthesis.	Partial (Profiles metric flaws but misses standardized quality tool implementation).	Standardization Void: Proposes no centralized clinical benchmark for verifying autonomous agentic behavior.
This Study	2017–2026 N = 178	Multi-era evaluation of biomedical summarization frameworks. PRISMA 2020 layout.	Architectural evolution, context decay, MQS/EAS errors, and CRL tiers. Micro-Technical & Mathematical Modeling.	Strictly bounded by field-wide publication biases and localized hospital "MIMIC dependency capture".	Yes (Direct Focus of RO1). It delivers a structured, longitudinal taxonomy mapping successive tech eras (2017–2026).	Yes (addresses Evaluation Fragmentation by enforcing a customized Methodological Quality Score [MQS] and 6-D EAS).	Yes (addresses the Standardization Void by establishing a Standardized 5-Level Clinical Readiness Level [CRL] pipeline).

In the last few years, several frameworks have been developed to assess clinical AI, including the Methodological Quality Score (MQS), the Evaluation Adequacy Score (EAS), and the Clinical Readiness Level (CRL). These instruments move beyond surface-level descriptive reviews by capturing the technical transparency, multidimensional validation depth, and translational maturity of Large Language Model (LLM) research [57,59,60]. Although they are methodologically ambitious, there have been questions about construct validity. These involve binary scoring in the MQS [61], the clinical rationale for equal weighting of the six EAS dimensions [62], limited sensitivity analysis [63], and unclear boundaries between readiness levels [64]. Addressing these points is crucial for aligning the pipeline with international reporting and evaluation standards such as Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence (TRIPOD-AI), Consolidated Standards of Reporting Trials–Artificial Intelligence (CON-SORT-AI), and Developmental and Collaborative Evaluation of Clinical AI (DECIDE-AI) [65,66]. Moreover, the frameworks show external validity, accounting for 64% of the variance in expert global-quality judgments, and offer preliminary insight into pathways for aligning technical benchmarks with clinical requirements [61,65,67].

To address these gaps, this review adopts three complementary assessment frameworks: the Methodological Quality Score (MQS), the Evaluation Adequacy Score (EAS), and the Clinical Readiness Level (CRL). MQS assesses reporting transparency and methodological completeness, EAS quantifies the comprehensiveness of evaluation practices, and CRL characterizes translational maturity across successive stages of clinical deployment [68–70]. Such separation is especially critical because good benchmark data can exist without good levels of validation, transparency, and safety issues [71,72]. Although EAS and CRL both incorporate clinically relevant evaluation concepts and acknowledge that they are related but not fully independent

constructs, they assess complementary layers of the translational pipeline, including evaluation completeness and translational maturity, respectively. Three Research Objectives (ROs) map directly onto the previous six gaps, consisting of two primary objectives and one secondary objective:

- RO1—Evaluation Audit: To evaluate the quality of reporting and evaluation rigor in 178 biomedical summarization studies systematically through the Methodological Quality Score (MQS) and Evaluation Adequacy Score (EAS) frameworks. This goal identifies strengths, weaknesses, and gaps in existing evaluation practices, including evaluations of factuality, safety, and expert validation (Gaps 3 and 4).
- RO2—Clinical Readiness Assessment: To assess the translational maturity of biomedical summarization systems using the five-level Clinical Readiness Level (CRL) metric and identify technical, operational, and governance barriers (Gap 5). It also assesses the empirical association between evaluation adequacy and clinical readiness, providing a structured audit of their interdependence using a quantitative multi-stage pipeline (Gap 6).
- RO3—Architectural Mapping: To produce a descriptive taxonomy of biomedical summarization architectures, such as encoder–decoder models, domain-adapted LLM models, and new agentic models. This analysis provides context for the field's development, but significant variation in tasks and datasets makes direct comparisons of performance across studies difficult (Gaps 1 and 2). Table 2 summarizes the research objectives, questions, and sub-questions.

Table 2: Alignment of research objectives and questions for the systematic review.

Research Objective (RO)/Focus	Research Question (RQ)	Sub-Research Questions (SRQ)
RO1: To systematically evaluate reporting quality and evaluation rigor in 178 biomedical summarization studies using the Methodological Quality Score (MQS) and Evaluation Adequacy Score (EAS) frameworks.	RQ1: How adequate are current evaluation frameworks for clinical safety, and what corpus gaps limit the generalizability of reported evaluation practices?	SRQ1.1: What is the distribution of evaluation metric types (lexical, semantic, factuality, human, LLM-as-judge) across the corpus, and what Evaluation Evaluation Adequacy Score (EAS) does each study achieve? SRQ1.2: What datasets are used for evaluation, and to what extent does dataset concentration (e.g., "MIMIC-dependency") constrain the generalizability and reproducibility of findings? SRQ1.3: What types of factual errors and hallucinations are reported, and what methods are utilized for their detection and quantification?
RO2: To assess the translational maturity of biomedical summarization systems using the five-level Clinical Readiness Level (CRL) metric and identify technical, operational, and governance barriers.	RQ2: At what level of clinical readiness does the field currently operate, what barriers prevent deployment, and is there a statistically significant association between evaluation adequacy and readiness?	SRQ2.1: What is the distribution of studies across the five Clinical Readiness Levels (CRL 1–5), and what proportion of the literature has progressed beyond laboratory benchmarking? SRQ2.2: To what extent do included studies address safety analysis, demographic bias, and regulatory compliance, and what implementation barriers are most frequently reported?

(Continued)

Table 2 (continued)

Research Objective (RO)/Focus	Research Question (RQ)	Sub-Research Questions (SRQ)
		SRQ2.3: Is there a statistically significant association between evaluation adequacy score (EAS) and clinical readiness level (CRL), and how does this association vary by architecture family or methodological quality?
RO3: To produce a descriptive taxonomy of Transformer- and LLM-based biomedical summarization architectures such as encoder–decoder, decoder-only, and agentic models to provide technological context for the field's evolution.	RQ3: How have architectures for biomedical summarization evolved, and what descriptive performance patterns are reported across model families and text types?	SRQ3.1: What are the dominant architectural paradigms and their longitudinal adoption trajectories across the three technological eras (2017–2026)? SRQ3.2: What are the reported performance values (ROUGE-L, BERTScore) stratified by architecture and text type, acknowledged as descriptive comparisons given high task heterogeneity? SRQ3.3: What adaptation strategies (supervised fine-tuning, RAG, agentic workflows) are employed, and what is the current state of prompt transparency and reproducibility?

Collectively, these objectives enable a multidimensional assessment of biomedical summarization research by integrating evaluation quality, clinical translation, and technological development within a unified systematic review framework. This study extends prior descriptive work by proposing a structured exploratory audit pipeline (MQS–EAS–CRL) to characterize reporting quality, evaluation adequacy, and clinical readiness within the analyzed corpus. The primary contributions of this systematic review are:

- Systematic Evaluation Audit: This study uses the Methodological Quality Score (MQS) and the Evaluation Adequacy Score (EAS) frameworks to quantitatively identify reporting quality and evaluation rigor from the literature. The analysis is based on various evaluation dimensions, such as factuality, safety, and expert validation, and includes a structured analysis of current evaluation practices and their limitations.
- Translational Readiness Assessment: This review introduces a five-level Clinical Readiness Level (CRL) framework to evaluate the maturity of biomedical summarization systems for real-world deployment. The framework identifies technical, operational, and governance challenges that influence clinical adoption and provides a structured pathway from laboratory evaluation to clinical implementation.
- Architectural Taxonomy: This review synthesizes evidence from 178 empirical studies published through early 2026 and develops a longitudinal taxonomy of biomedical summarization architectures. It describes the field's evolution through specialized encoder–decoder systems, to domain-adapted LLMs, and more recently, to agentic systems, and presents an overview of the current research landscape.
- Given the substantial heterogeneity of datasets, clinical domains, evaluation protocols, and reporting practices across the included studies, all comparative analyses presented in this review should be interpreted descriptively rather than as formal comparative effectiveness evaluations.

2 Materials and Methods

This study was conducted as a systematic review of Transformer- and LLM-based biomedical text summarization architectures, using a structured evidence synthesis methodology to characterize architectural evolution, evaluation adequacy, and clinical readiness. This systematic review was conducted in accordance

with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 reporting guideline and the Population–Concept–Context (PCC) framework, which is recommended for evidence synthesis in emerging technological domains.

- Population: Biomedical and clinical text corpora, such as electronic health records (EHRs), discharge summaries, radiology and pathology reports, clinical notes, and biomedical literature.
- Concept: Transformer-based models, large language models (LLMs), domain-adapted foundation models, and multi-agent systems developed for biomedical text summarization and related information synthesis tasks.
- Context: Experimental evaluations, institutional validation studies, benchmarking environments, and real-world clinical deployment settings.

The study protocol was preregistered on the Open Science Framework (OSF Registration ID: [4zrs7]). The review protocol was structured around three master Research Objectives (ROs) designed to address high-priority gaps in the field of Large Language Model (LLM)-based biomedical summarization.

2.1 Literature Search Strategy

Five major bibliographic databases (PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore, and ACM Digital Library) were searched systematically. ACL Anthology, arXiv, bioRxiv, and medRxiv were used as supplementary sources to capture emerging trends in generative AI and agentic workflows. Studies published between 01 January 2017 and 22 March 2026, were considered for inclusion. The search period was chosen to cover the evolution of biomedical summarization systems from the emergence of Transformer-based architectures to recent LLM-driven and agentic approaches. The same timeframe was used across all databases to ensure uniformity in study identification data.

The search was restricted to original empirical studies published in English. The unified search strategy is summarized in Table 3, with the full details of the search query provided in Appendix A (Table A1). The search strategy employed three main concept blocks with the use of the Boolean operators:

- Block A (Models): Transformer architectures and large language models, including terms such as Transformer, BERT, GPT-4, Llama, and foundation model.
- Block B (Task): Biomedical text summarization tasks, including summarization, abstractive summarization, and impression generation.
- Block C (Context): Biomedical and clinical domains, including radiology, electronic health records, discharge summaries, and pathology.

Table 3: Summary of the unified search strategy and quantitative results across bibliographic databases.

Databases Searched	Unified Search Query (Conceptual Structure)	Filters Applied	Records Retrieved
PubMed, Scopus, WoS, IEEE, ACM, ACL Anthology	(Block A: “summarization” OR “abstractive” OR “impression generation” OR “BTS”) AND (Block B: “biomedical” OR “clinical” OR “EHR” OR “medical text”) AND (Block C: “transformer” OR “LLM” OR “GPT-4” OR “BERT” OR “attention” OR “conversational agents”)	Peer-reviewed articles; Full conference papers; 01 January 2017 to 22 March 2026; English language.	2130

2.2 Eligibility Criteria

Studies were included if they fulfilled all the following Inclusion Criteria:

- (I1) Study Design: Original empirical research that appears in peer-reviewed journal articles, complete conference proceedings, or verified academic preprints.
- (I2) Model Architecture: Explicit use of at least one Transformer-based Sequence to Sequence model architecture or Large Language Model architecture, such as Encoder-only, Encoder-Decoder, Decoder-only, Hybrid or Agentic (Multi-turn) models.
- (I3) Core natural language processing (NLP) Task: Automatic generation of condensed textual representations from source text documents, which can be extractive, abstractive or Retrieval-Augmented Generation (RAG) summarization pipelines.
- (I4) Clinical Domain: Application to biomedical or healthcare text sources, such as Electronic Health Records (EHRs), discharge summaries, radiology reports, pathology reports, clinical notes, or peer-reviewed biomedical literature.
- (I5) Evaluation Protocol: Reporting of at least one replicable quantitative or expert-based evaluation measure, including surface lexical overlap (ROUGE, BLEU), advanced semantic alignment (BERTScore), or structured human clinical assessment.
- (I6) Language and Timeframe: Original English-language studies published or formally archived between 01 January 2017 and 22 March 2026.

On the other hand, studies were systematically discarded when they met any of the Exclusion Criteria:

- (E1) Publication Type: Systematic reviews, scoping overviews, editorials, commentaries, letters, book chapters, or non-empirical reports.
- (E2) Task Misalignment: Studies that examined language processing were only on non-summarization tasks (e.g., named entity recognition (NER), relation extraction, text classification labeling, without including text generation tasks.
- (E3) Algorithmic Obsolescence: Papers that make use of pre-Transformer or non-attentional machine learning methods such as rule-based systems, TF–IDF, LexRank, Word2Vec, or classic recurrent neural network architectures, that do not apply attention approaches.
- (E4) No Empirical Validation: Failed to obtain quantitative/qualitative experimental validation tests for a theoretical or architectural proposal.
- (E5) Non-English Publications: Full-text articles in languages other than English. The linguistic limitation is explicitly recognized as a field-wide vulnerability in the Limitations section.
- (E6) Incomplete Reports: Conference abstracts, poster summaries, slide presentations or studies with a non-accessible full-text companion manuscript.

Studies based exclusively on pre-Transformer architectures were excluded because the review focused on Transformer- and LLM-based biomedical summarization systems. This restriction ensured methodological consistency within the analyzed corpus. Consequently, this scoping boundary reflects an objective, chronological technological shift across the artificial intelligence landscape rather than a post hoc filter optimization, thereby preserving the internal validity and association capacity of all downstream statistical and regression models. Preprints from the ACL Anthology, arXiv, bioRxiv, and medRxiv were reviewed using the same criteria as peer-reviewed studies, where the available version was kept.

2.3 Selection and Screening

To reduce selection bias, study screening was conducted independently by two reviewers (trained research assistants). Titles and abstracts were then assessed for eligibility against the pre-defined criteria after the duplicates were removed. Potentially eligible studies were then subjected to full-text review by the same reviewers. In cases of disagreement, the matter was resolved through discussion and consensus; if a tie persisted, the author (acting as the senior adjudicator) was consulted for a final decision. To evaluate consistency in the selection of studies and data extraction, inter-rater reliability was tested. Cohen's kappa (κ) was used for decisions on screening of titles and abstracts. Agreement was determined using the Intraclass Correlation Coefficient (ICC) based on the two-way mixed-effects absolute-agreement model for the following assessments: Methodological Quality Score (MQS), Evaluation Adequacy Score (EAS), and Clinical Readiness Level (CRL).

2.4 Operational System Definitions

For the purposes of this review, Fig. 1 illustrates the general workflow of a Biomedical Summarization system that converts clinical text into a compact summary for subsequent use in the healthcare domain.

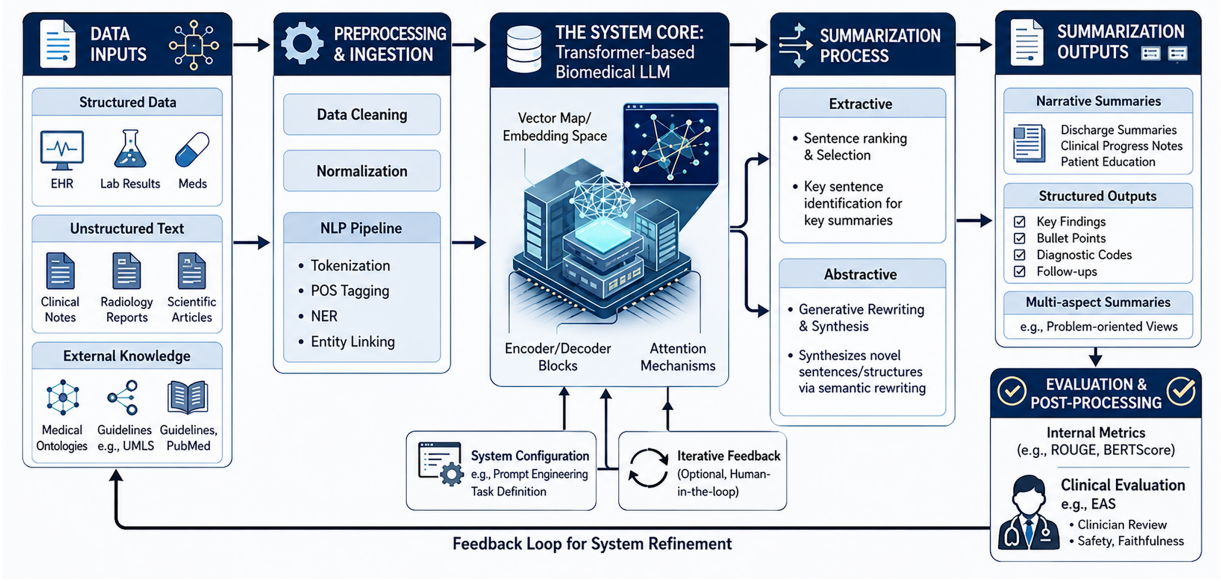

Figure 1: Baseline operational workflow of a generic biomedical summarization system.

This explicitly visualizes multi-source data inputs, data preprocessing pipelines, core model spaces as provided by Transformer, and diverse clinical outputs, thereby establishing a baseline for subsequent analysis of architectural and readiness stages. A biomedical summarization system is a pipeline that uses statistical, neural, or large language models to summarize medical domain text. System inputs include unstructured or semi-structured biomedical texts, like electronic health records (EHRs), clinical notes, discharge summaries, nursing handovers, intensive care unit reports, radiology and pathology reports, and biomedical literature. System outputs are concise summaries designed to synthesize information relevant to the clinical setting. These outputs can be used for clinical decision making, patient documentation, administrative processes, or patient communication. Systems designed exclusively for classification, entity extraction without synthesis, or single-word labeling were excluded under this definition.

This study aims to differentiate the summarization approaches using conventional LLM architectures from those using newer agentic architectures. It defines an agentic summarization system as one that uses an LLM as part of an autonomous multi-step process rather than as a single-question generation system. A system must have at least two of the following capabilities to be agentic:

- Planning: Decomposition of complex summarization tasks into multiple steps or decision processes.
- Memory: Maintenance of contextual information or execution states across multiple interactions.
- Tool Use: Integration with external tools, databases, application programming interfaces (APIs), or biomedical knowledge resources.
- Self-Reflection: Internal verification, critique, or refinement mechanisms used to evaluate and improve generated outputs.

These operational definitions were used a priori throughout the extraction and classification of the data into architectures.

2.5 Measurement Pipeline: MQS-EAS-CRL Framework

Current assessment methods, such as RoB 2, ROBINS-I, TRIPOD-AI, and PROBAST-AI, were mainly designed for clinical trials and prediction models, and do not adequately reflect the methodological and evaluation features of natural language processing (NLP) and large language model (LLM) studies [48,61,65]. LLM studies, in particular, are frequently evaluated using benchmarks, with little to no clinical outcomes over time and varied validation methods. The absence of standardized assessment criteria for methodological rigor, evaluation completeness, and translational preparedness of healthcare LLM research has also been documented in recent reviews [53,70,73,74].

To overcome these limitations, the integrated MQS–EAS–CRL measurement pipeline was implemented to assess the Methodological Quality Score (MQS), the Evaluation Adequacy Score (EAS), and the Clinical Readiness Level (CRL). These Assessment frameworks collectively analyze the studies across three complementary analytical layers, including methodological quality and reporting transparency (MQS), evaluation completeness and the depth of data validation, and translational readiness for clinical deployment.

The MQS pipeline evaluates methodological rigor, transparency, and reproducibility. The EAS measurement pipeline considers six aspects of validation to assess its comprehensiveness: lexical, semantic, factuality, human, clinical, and safety. The CRL analytical model defines the maturity levels of a biomedical summarization system along a continuum from laboratory research to clinical use. A summary of the alignment of these audit pipelines with existing standards is provided in Table 4, which includes TRIPOD-AI, CONSORT-AI, DECIDE-AI, The Evaluation Framework for Health Information Technology (TEHAI), and Technology Readiness Levels (TRL).

Table 4: Mapping the MQS–EAS–CRL infrastructure to established clinical AI reporting standards.

Pipeline Elements	Core Construct	Established Standards Aligned with	Key Design Principle
MQS	Methodological rigor, transparency, and reproducibility	TRIPOD-AI, CONSORT-AI, PRISMA 2020, PROBAST-AI	Checklist-based binary items capture presence/absence of transparency indicators [32,75,76].

(Continued)

Table 4 (continued)

Pipeline Elements	Core Construct	Established Standards Aligned with	Key Design Principle
EAS	Multi-dimensional evaluation completeness	TRIPOD-LLM, DECIDE-AI, TEHAI, multi-criteria NLP evaluation frameworks	Six dimensions (lexical, semantic, factuality, human, clinical, safety) reflect the full evaluation spectrum required for clinical AI [37,67,77,78].
CRL	Translational and deployment maturity	Technology Readiness Levels (TRL), FDA AI/ML lifecycle, DECIDE-AI staged evaluation, TEHAI	Five ordered levels mirror clinical AI lifecycle from bench research to sustained deployment [79].

These three instruments are combined to enable a structured process for identifying the methodological, maturity-level, and clinical readiness of the selected corpus. It also allows for the identification of asymmetric maturity; studies have demonstrated high methodological quality but have been limited in clinical readiness due to a lack of validation, safety evaluation, or real-world evaluation. The MQS–EAS–CRL Scoring system facilitates an all-encompassing review of methodological quality, evaluation adequacy, and translational maturity, and ensures that the review follows current best practices for trustworthy clinical AI [57,79,80].

2.5.1 Methodological Quality Score (MQS)

The MQS is an 8-item binary checklist developed based on clinical AI reporting frameworks, including Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis–Artificial Intelligence (TRIPOD-AI), Consolidated Standards of Reporting Trials–Artificial Intelligence (CONSORT-AI), Checklist for Artificial Intelligence in Medical Imaging (CLAIM), and Prediction model Risk Of Bias Assessment Tool–Artificial Intelligence (PROBAST-AI) [32,76,81]. It is adopted in this study to assess the methodological transparency and reproducibility of corpus studies, using four main domains: Model Transparency (MT), Data Quality (DQ), Reproducibility (RP), and Evaluation Rigor (ER) [82]. A binary pass-fail scoring rubric was adopted to minimize subjective grading variance and enhance inter-rater consistency, auditability, and feasibility across 178 heterogeneous studies [65]. MQS is not used as an eligibility criterion.

MQS is used to categorize the corpus studies into three distinct quality tiers, High Quality (MQS 7–8), Moderate Quality (MQS 4–6), and Low Quality (MQS 0–3) This classification is designed to identify systemic bottlenecks, such as the reproducibility gap between high-performing proprietary models and verifiable open-source architectures. The full MQS scoring manual is provided in Appendix B (Table A2). All included studies were scored based on the MQS (0–8 scale). The scoring rubric was informed by current reporting recommendations for AI models, which emphasize data provenance, model description, and the availability of open-science artifacts.

2.5.2 Evaluation Adequacy Score (EAS)

Existing LLM-based biomedical summarization studies predominantly rely on lexical and accuracy-based metrics for evaluation, with comparatively limited attention given to factuality, safety, human

assessment, and clinical validity [77,78]. As a result, strong benchmark performance may not necessarily reflect clinical reliability or real-world utility [71,77,83]. To address this limitation, the Evaluation Adequacy Score (EAS) was adopted in this review to assess clinical relevance of evaluation practices in biomedical summarization research [37,78]. The EAS was informed by recent clinical-AI evaluation standards summarized in Table 4, and customized here to explore whether studies report evaluation across dimensions commonly associated with trustworthy clinical AI.

The EAS assessment model comprises six dimensions: Lexical Evaluation (D1), Semantic Evaluation (D2), Factuality Assessment (D3), Human Evaluation (D4), Clinical Validity (D5), and Safety and Bias Assessment (D6). These dimensions are used to convey both technical performance and clinically meaningful evaluation outcomes. Each dimension is scored on a three-point ordinal scale: 0 (not assessed), 1 (partial or qualitative assessment), and 2 (comprehensive or systematic assessment). The EAS score is calculated by summing the six-dimensional scores, with a higher score indicating more extensive and clinically appropriate evaluation practices. The EAS total score is computed as:

$$EAS_{(raw)} = \sum_{i=1}^{6} di \tag{1}$$

$$EAS_{norm} = EAS_{(raw)}/12 \tag{2}$$

where *di* represents the score assigned to each evaluation dimension (0, 1, 2). The raw *EAS* score ranges from 0 to 12 and is subsequently normalized to the interval [0, 1] using Eqs. (1) and (2).

Predefined operational criteria were developed for each dimension of EAS to ensure consistency in scoring and minimize assessor subjectivity. To assess whether the six ordinal EAS dimensions collectively measured the underlying latent construct of evaluation adequacy, Cronbach's alpha (α) was calculated across all dimensions. Spearman correlation analysis was used to further investigate the relationship between methodological quality and evaluation adequacy by correlating the scores of the MQS and EAS [61,83,84]. The full scoring decision rules for each level (0, 1, or 2) for each dimension are included in Appendix B (Table A3). In D4 (Human Evaluation), scores were awarded only when the summary outputs were evaluated directly by qualified human domain experts, such as physicians, clinical specialists, or senior healthcare practitioners. LLM-as-a-judge frameworks, self-evaluation mechanisms, and multi-agent scoring systems were not included as evidence of human evaluation and were not awarded credit for D4.

In addition to the scored dimensions, four contextual descriptors were extracted to qualify the EAS: (i) use of hallucination taxonomies, (ii) adoption of LLM-as-judge evaluation, (iii) reporting inter-annotator agreement, and (iv) prompt transparency. MQS and EAS scores were used only for analytical stratification and were not used as eligibility criteria for the study. Scores served as stratification variables, enabling statistical differentiation between foundational benchmarks and translationally mature systems. These descriptors serve as additional context for understanding evaluation practices. To ensure clinical safety and relevance across diverse subdomains, the Safety Audit is executed dynamically against a structured Task-Specific Risk Matrix (see Appendix B, Table A4). While the EAS and CRL scoring models evaluate different aspects of the translational pipeline, they are related but partially overlapping constructs. Dimensions such as human evaluation (D4), clinical validity (D5), and safety assessment (D6) in the EAS share the same underlying scoring logic as the higher CRL stages.

2.5.3 Clinical Readiness Level (CRL): Staging Clinical Maturity

Clinical readiness, real-world integration, and deployment maturity are highlighted in recent evaluation frameworks, including The Evaluation Framework for Health Information Technology (TEHAI), AI for

IMPACTS, and other implementation-focused frameworks [79]. This review builds on these principles by using the Clinical Readiness Level (CRL), an ordinal scale adapted from NASA's Technology Readiness Levels (TRLs) for the translational maturity of biomedical summarization systems [85–87].

The CRL is a five-level ordinal scale that categorizes the Clinical maturity of biomedical summarization systems, from foundational benchmarking (CRL-1) to routine clinical integration (CRL-5) [88,89]. The CRL framework was utilized to classify studies according to three criteria: (i) data provenance (public benchmarks or institutional clinical data); (ii) the extent of human expert involvement; and (iii) the degree of workflow integration within clinical practice.

- CRL-1 (Benchmarking): Automated evaluation on public benchmark datasets, without human expert review.
- CRL-2 (Expert Review): Public or retrospective data with informal or limited expert evaluation.
- CRL-3 (Institutional): Structured evaluation using institutional or private clinical data with multi-rater expert assessment.
- CRL-4 (Prospective): Prospective validation within a real or simulated clinical workflow.
- CRL-5 (Operational): Sustained routine clinical deployment supported by operational monitoring and error-reporting mechanisms.

The highest CRL stage reported with validation evidence was used for each study. Predefined criteria were used in all studies to ensure consistent classification. For instance, the transition from CRL-2 to CRL-3 was based on evaluation of real-world institutional data and structured clinical expert review, while the transition from CRL-3 to CRL-4 was based on evidence of prospective clinical implementation. CRL stages characterize the translational progression of biomedical summarization systems from laboratory development to routine clinical deployment and are not intended to quantify the rigor of evaluation or methodological quality. The full CRL staging criteria and decision rules are given in Appendix B (Table A5).

Furthermore, seven clinical maturity indicators were extracted for each CRL score, including external validation, regulatory reporting, and clinician-in-the-loop involvement, to provide additional translational context. In addition, one barrier statement was taken from the Discussion or Limitations section of each study and categorized into six pre-defined domains: Technical, Accuracy, Operational, Ethical, Data Scarcity, and Regulatory. This qualitative analysis was used to enhance the ordinal CRL staging by highlighting the common barriers that may hinder progression toward clinical deployment. The integrated MQS–EAS–CRL Methodological structure was developed to provide complementary perspectives on reporting quality, evaluation adequacy, and Clinical maturity. Relationships among these dimensions were investigated descriptively using correlation and regression analyses to characterize empirical associations within the corpus rather than to establish causal dependence.

2.5.4 Framework Validation Strategy

The proposed audit pipeline for the MQS, EAS, and CRL was validated through a multi-stage process to ensure its construct validity, reliability, robustness, and utility for association. The validation protocol aims to ensure that the Assessment approach provides repeatable, statistically sound methodological quality, evaluation adequacy, and translation-readiness assessment for biomedical summarization research.

- **Construct validity:** To establish the construct validity, the domains of each Scoring system were mapped to the recognized reporting and evaluation standards as described previously in Table 4. This alignment demonstrates that the frameworks interpret and apply existing methodological constructs rather than arbitrary scoring criteria. Framework operational mechanics were strategically optimized

to maximize objective auditability and neutralize potential biases. To minimize subjective grading variance, maximize inter-rater reproducibility, and ensure consistency with checklist-based methodological assessment tools, binary (0/1) scoring was used for MQS. To prevent the introduction of a preconceptual prioritization bias in the absence of universal consensus in clinical practice, equal weighting across the six dimensions of the EAS was adopted as a conservative baseline.

- **Reliability and Internal Consistency:** Screening, data extraction, and instrument scoring were performed independently by two trained and compensated research assistants, using the predesigned scoring manuals. The assistants underwent standardized training prior to study selection and data extraction. Disagreements identified after independent coding were adjudicated by the author, who remained blinded to the calibration exercise and did not participate in the initial coding process. The inter-rater reliability (IRR) of MQS, EAS, and CRL was evaluated using a two-way mixed-effects ICC (absolute agreement model) (ICC). Cohen's κ was used to assess item-level agreement on binary MQS indicators, while weighted κ was used to assess item-level agreement on CRL assignments that have an ordinal level. The reliability coefficients were interpreted according to methodological recommendations, with ICC ≥ 0.70 considered acceptable [90,91]. The internal consistency of MQS and EAS was checked separately. For the eight binary MQS items, the Kuder–Richardson 20 (KR-20) coefficient was computed, while Cronbach's alpha (α) was computed across the six ordinal EAS dimensions to determine if they collectively measured the latent construct of evaluation adequacy. To address the sparsity of high-level CRL samples (CRL 4–5), the primary inferential focus is placed on a binary stratification (CRL 1–2 vs. CRL ≥ 3), treating higher levels as descriptive supplementary evidence.
- **Discriminative validity:** Discriminative validity was assessed by comparing MQS and EAS scores across studies with varying levels of Clinical maturity. Studies were categorized as early-stage systems (CRL 1–2) or clinically advanced systems (CRL 3–5), and differences in MQS and EAS scores were evaluated by the Mann–Whitney U test. The purpose of this analysis was to establish if there was an increasing systematic improvement in methodological rigor and evaluation completeness as the translation process advanced. The relationship between MQS, EAS, and CRL was explored by Spearman's rank correlation coefficient (ρ). Correlation analyses were performed for the overall corpus and stratified by architecture family to investigate whether evaluation adequacy was more strongly associated with clinical readiness than methodological reporting quality alone. A Receiver Operating Characteristic (ROC) analysis was conducted with the minimum level of evaluation adequacy associated with institutional feasibility as the positive outcome. The EAS corpus-derived exploratory reference point was determined by the Youden Index. A bootstrap sample (10,000 replications) was then performed to estimate 95% confidence intervals and to evaluate the robustness of the exploratory reference point.

Although EAS and CRL evaluate distinct conceptual dimensions, they should be regarded as complementary yet partially overlapping analytical models, as several EAS dimensions (particularly expert evaluation, clinical validation, and safety assessment) align conceptually with higher stages of clinical readiness. Consequently, subsequent statistical analyses are interpreted as characterizing empirical associations rather than relationships between fully independent constructs.

2.5.5 Sensitivity Analyses, Comparative Testing, and Inferential Modeling

Multiple complementary analyses were performed to assess the robustness, generalizability, and validity of the proposed MQS–EAS–CRL framework. The reliability of the Evaluation Adequacy Score (EAS) was assessed by assigning greater weight to the Factuality Assessment (D3), Clinical Validity (D5), and Safety Assessment (D6). Spearman correlation analysis was used to characterize monotonic associations among MQS, EAS, CRL, and performance metrics without implying causal relationships among the evaluated

constructs. Consistent with the high heterogeneity of the corpus, architectural performance comparisons are interpreted as descriptive patterns rather than robust inferential rankings. To assess the possible impact of publication status and of aggregating studies by dataset and research group to account for non-independence of observations, additional sensitivity analyses were carried out after excluding non-peer-reviewed preprints and after combining studies from the same dataset and research group. The analyses focused on whether the main relationships observed in the MQS–EAS–CRL framework were consistent across different corpus types.

Comparative inferential analyses were conducted to determine whether methodological quality, evaluation adequacy, and Clinical maturity differed across architectural paradigms and technological eras. Chi-square tests of independence and/or Fisher's exact tests were used to investigate relationships between architecture families (Encoder–Decoder, Decoder-only, and Hybrid/Agentic systems) and evaluation practices where expected cell frequencies were low. The Kruskal–Wallis H test was used to analyze differences in the MQS, EAS, and CRL scores across the three technological epochs.

A multivariable ordinal logistic regression model was created, with CRL (Levels 1–5) specified as the ordinal dependent variable. Independent variables were MQS, EAS, publication period, architecture family, and data set type (public vs. institutional/private). This model was used to estimate the relative association of reporting quality and evaluation adequacy with clinical readiness. The regression framework characterizes statistical associations rather than independent causal relationships, acknowledging partial conceptual overlap between EAS dimensions (D4–D6) and higher CRL stages that involve clinical validation. All statistical analyses were performed using Python 3.12. Statistical significance was established a priori at a two-sided α level of 0.05. Where appropriate, effect sizes and 95% confidence intervals were reported to facilitate interpretation of practical significance.

2.5.6 Reviewer Training and Calibration

The MQS, EAS, and CRL were piloted separately by the same two reviewers on 10 randomly selected studies prior to full-corpus scoring. This calibration exercise helped clarify operational definitions, establish a common interpretation of scoring, and address potential ambiguities in applying the framework. Specific focus was given to boundary cases in the CRL rubric, particularly the distinction between informally reported author observations and the structured expert-evaluation evidence used to assign higher CRL levels in the scoring scheme. After calibration, the two reviewers separately scored the entire study corpus (N = 178) with standardized scoring manuals. Consensus scores were used as the final dataset, and discrepancies were resolved through structured consensus discussions with the author.

2.6 Data Extraction and Synthesis

A standardized data extraction form was developed a priori to collect variables that support Research Objectives RO1-RO3 and to fill in the Methodological Quality Score (MQS), Evaluation Adequacy Score (EAS), and Clinical Readiness Level (CRL) frameworks. The data were extracted at the individual study level and sorted into the seven predefined domains as shown in Table 5. The extraction instrument was pre-tested to guarantee uniformity, comprehensiveness, and consistency with the review goals.

Sample inflation and double counting in longitudinal and comparative analyses were avoided by only including the main model suggested, optimized, or emphasized by the study authors. Secondary baselines, comparator systems, and benchmark models were recorded descriptively but were not included as independent analytical units. The methodical approach allowed for a consistent methodology and provided the interpretability of the temporal and architectural trends throughout the corpus. For studies with borderline or ambiguous evidence, coding was conservative, and borderline studies were assigned to the lower level

to limit the potential for overrating methodological quality, evaluation adequacy, or translation readiness. Hallucination and Clinical Failure-Mode Coding were analyzed using two schemas: a broad reporting group (n = 99), comprising studies reporting any hallucination or clinical failure mode, and an explicit taxonomic auditing subset (n = 78), comprising studies using structured multi-label hallucination taxonomies. Multi-label coding was permitted; therefore, category frequencies exceed 100%. The broader group was used for prevalence analyses, whereas the taxonomic subset was used for structured auditing analyses. Detailed procedures for data extraction, calibration protocols, and decision rules are included in the Appendix B.

Table 5: Structured data extraction domains and research objective alignment.

Domain	Extracted Variables	Objective Alignment
Metadata	Bibliographic and contextual details, including author, publication year, venue type, country, funding source, and open access status.	Descriptive Baseline
MQS scoring	Binary pass/fail (0/1) for each of the eight MQS items across the four domains (MT, DQ, RP, ER); MQS = sum of items, range 0–8.	RO1 (Methodological Quality)
(EAS) Scoring	Ordinal depth score (0 = absent, 1 = partial/qualitative, 2 = comprehensive/systematic) for each of the six EAS dimensions (D1–D6); EAS = raw sum/12, yielding a normalized score in [0, 1], where 0 indicates no evaluation dimension addressed and 1 indicates all six dimensions comprehensively covered.	RO1 (Evaluation Adequacy)
CRL and deployment context	RL level assignment (1–5) based on data provenance and expert involvement; deployment architecture, and reported technical/operational/governance barriers. Includes clinician-in-the-loop (CiTL) and regulatory compliance.	RO2 (Clinical Readiness)
Architecture and adaptation	Architectural family (encoder-decoder, decoder-only, hybrid/agentic); adaptation strategies; and prompt engineering methods.	RO3 (Taxonomy)
Summarization paradigm	(Extractive, abstractive, or hybrid), prompt engineering strategy (zero-shot, few-shot, chain-of-thought), input text type, dataset identity, size, and language.	RO3 (Corpus Mapping)
Performance Metrics	Numerical values for ROUGE-L, BERTScore, and domain-specific metrics. These are strictly coded as descriptive performance patterns to provide technical context. Direct inferential rankings are avoided to account for high task and dataset heterogeneity.	RO3 (Descriptive Taxonomy Context)

2.6.1 Data Synthesis and Quantitative Analysis

A convergent mixed-methods synthesis approach was used to connect the quantitative evidence to the MQS–EAS–CRL assessment framework. No formal meta-analysis was conducted because there

was considerable heterogeneity across the summarization tasks, clinical domains of the studies, datasets, evaluation protocols, and reporting metrics, and thus, the assumption of a common underlying effect size that would allow pooled effect estimation was not met. Instead, evidence synthesis combined structured descriptive analysis, quantitative performance synthesis, and framework-based validation analyses.

Several normalization procedures were conducted to ensure the homogeneous treatment of computational and methodological data. First, the studies used several models (GPT-4, Llama-3, etc., or other competing architectures) and one of the main models (the model proposed, optimized, or mentioned by the study authors) was chosen as the unit of analysis. Comparator and baseline models were kept as descriptions, but because they were not analyzed as separate observations, they did not inflate the sample size or double-count in trend analyses. Second, reported metrics such as ROUGE-L, BERTScore, and other evaluation metrics were standardized to the common range [0, 1] when this was mathematically possible to allow comparison of different studies. Third, variables that could not be extracted from the original publications were coded as not reported (NR) rather than inferred or imputed.

Observations where a metric was missing were not included in quantitative summaries but were included in descriptive and categorical summaries. The study characteristics, methodological attributes, evaluation practices, and maturity indicators for the translation were summarized using descriptive statistics. For categorical variables, frequencies and percentages are reported. The non-normal distributions observed across datasets and evaluation settings, and the significant differences in heterogeneity, are addressed by summarizing continuous and ordinal variables by their medians and interquartile ranges (IQR). To compare biomedical summarization approaches, performance outcomes were further aggregated to the categories of architecture family, summarization paradigm, and source text type.

2.6.2 Risk of Bias Assessment

Due to a lack of a validated risk-of-bias instrument designed specifically for LLM- and NLP-based biomedical text summarization research, risk of bias was evaluated qualitatively at the corpus level. Eight pre-defined binary indicators were used to assess issues of interpretation, generalizability, and translational relevance of the evidence base:

- MIMIC-Ecosystem Dependency: Use of MIMIC-III, IV, or Chest X-Ray (CXR) datasets.
- Radiology-Only Task Focus: Limiting the research to radiology reports or imaging impressions.
- English-Language Concentration: Exclusive use of English-language corpora.
- Academic Preprint Status: Studies retrieved from arXiv, bioRxiv, or medRxiv, and not published yet.
- Closed-Weight Proprietary Models: Use of models (e.g., GPT-4, Claude) via API without weight access.
- Unavailability of Source Code: Lack of publicly available code repositories to facilitate reproducibility
- Unavailability of Exact Prompts: Failure to share the specific textual instructions used.
- Institutional/Private Data Dependence: Dependence on non-public or restricted clinical data.

These indicators were assigned binary flags for each study (N = 178) to assess whether the association between the Evaluation Adequacy Score (EAS) and Clinical Readiness Level (CRL) was confounded by dataset provenance, clinical domain, or publication status.

To account for potential sources of heterogeneity, eight study-level risk-of-bias indicators were extracted from the master dataset, including MIMIC dependency, radiology focus, English-language concentration, preprint status, closed-weight/API model use, source code unavailability, prompt configuration unavailability, and institutional/private data use. To enhance the robustness of statistical inference and provide thorough control for confounding, all audit indicators were entered as independent variables into the multivariable ordinal logistic regression analysis and stratified sensitivity analyses. In particular, each indicator was

analyzed as a covariate in the CRL model, and the association between EAS and CRL was re-assessed in strata defined by the presence and absence of each indicator. Confidence intervals were obtained through 10,000 bootstrap iterations, and Fisher's r-to-z transformations with Holm-Bonferroni corrections were used to evaluate differences between strata for multiple comparisons. These analyses together evaluated the robustness of the observed EAS–CRL relationship to potential sources of bias and heterogeneity within the corpus.

2.7 Ethical Considerations and AI Use Disclosure

This systematic review was performed without any human subjects, patient data, clinical interventions, or data collection. Therefore, there was no need for institutional ethical approval or informed consent. The review process from study identification to screening, eligible study selection, data extraction, quality appraisal, framework development, statistical analysis, and evidence synthesis was all performed manually, ensuring methodological rigor, transparency, and reproducibility.

Generative artificial intelligence (AI) tools were used solely for limited language refinement to improve readability and were not involved in the literature search, study selection, data extraction, scoring, statistical analyses, result interpretation, scientific reasoning, conclusion generation, or any scientific decision-making. AI-assisted image generation using NanoBanana was employed only for the initial conceptual drafting of selected graphical illustrations. Multiple prompt refinements and manual revisions were performed to develop the final figures. All AI-assisted figures were critically reviewed, manually edited, and scientifically verified by the author, including all workflows, processes, labels, symbols, annotations, relationships, and graphical elements, to ensure that they accurately represented the underlying methodology and findings. The final structure, scientific content, and presentation of every figure were determined exclusively by the author. All manuscript text and figures underwent comprehensive human review to ensure scientific accuracy, internal consistency, reproducibility, and compliance with the journal's reporting guidelines. The author take full responsibility for the accuracy, integrity, and scientific validity of all content in this manuscript.

3 Results

3.1 Study Characteristics

A total of 2130 records were initially identified through systematic searches of the selected databases. Following the elimination of 471 duplicates, 1659 unique records were screened against the predefined PCC criteria based on title and abstract. 1302 records were excluded during the first screening phase; 357 articles were assessed at full text. Another 179 studies were eliminated because they lacked empirical validation or lacked quantitative evaluation, or utilized non-transformer architectures, non-summarization tasks, or insufficient methodological description. 178 studies that passed all eligibility criteria were included in the final evidence synthesis. The complete study selection process is presented in the PRISMA 2020 flow diagram (Fig. 2).

3.1.1 Architectural and Geographic Distribution

The publication activity was divided into three technological epochs, corresponding to the basic architectural changes.

- Era 1 (Transformer Foundations, 2017–2020): Represented the earliest adoption of attentional architectures within the corpus, excluding legacy non-attentional methods. This was the least represented category, accounting for 1.2% (n = 2) of the total corpus.

- Era 2 (Seq2Seq Maturity, 2021–2023): represented by 23% (n = 41) of the included studies, introduced domain-specific encoder-decoder models such as BART, T5, and PEGASUS.
- Era 3 (Agentic Orchestration, 2024–2026): This is the dominant era represented in the corpus, with 75.8% (n = 135) of studies emerging during this epoch.

This concentration identifies a shift in architectural adoption between 2023 and 2024, in which the field transitioned from task-specific text compression toward large-scale generative reasoning and multi-agent clinical workflows.

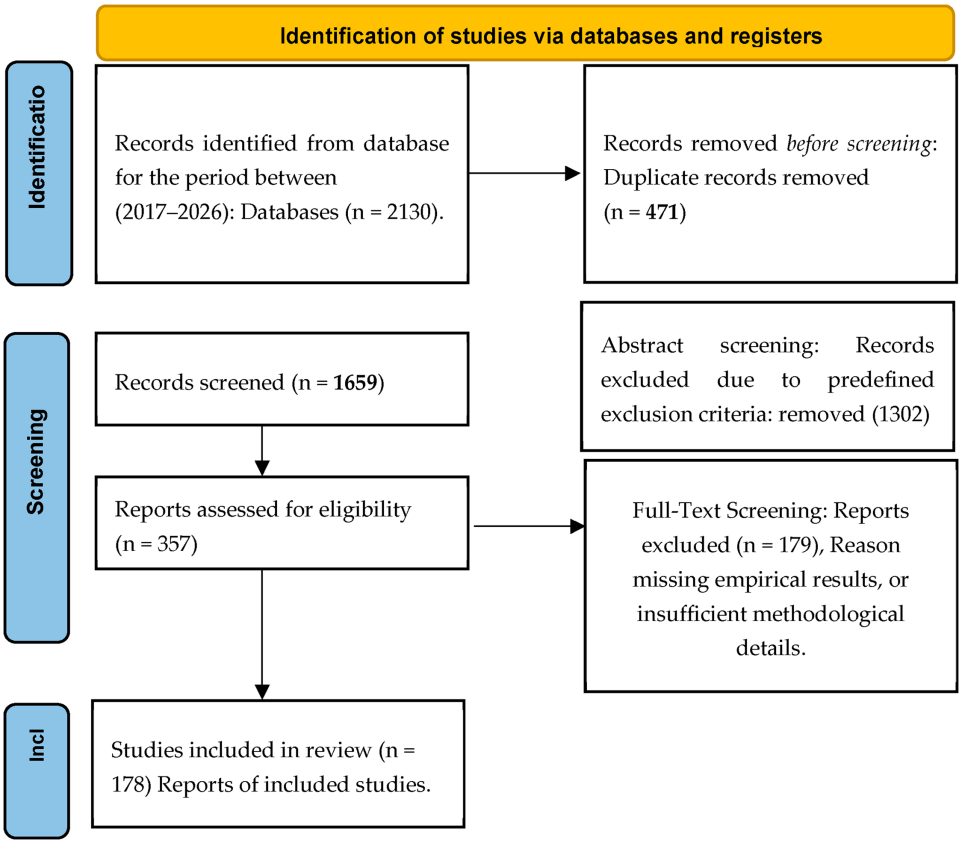

Figure 2: PRISMA 2020 flow diagram illustrating the systematic study selection process.

The United States (n = 42, 23.6%) and China (n = 37, 20.8%) led biomedical-summarization research output (Fig. 3A). Other contributions from the UK, Australia, India, Canada, Germany, and some other countries show the increasing international reliance on foundation models and agentic methods for clinical text summarization. Even with this geographical variation, 147 studies (82.6%) used English-language datasets, mostly from the MIMIC ecosystem, with the remaining 16.9% comprising regional studies. This concentration in English-language environments is partly attributable to the review's eligibility criteria, which restricted inclusion to English-language publications. Moreover, 75.8% of the corpus was published in the past two years, indicating the growing rate of LLM-based biomedical summarization, but also creating a temporal bias since many systems have not yet been independently replicated or long-term clinically validated. This suggests that the evidence is geographically, linguistically, and temporally limited and may not be generalizable across diverse healthcare settings for reported performance and readiness results.

Fig. 3B shows that 64.0% (n = 114) of the included studies were published in peer-reviewed journals, 27.0% (n = 48) in conference proceedings, and 9.0% (n = 16) as preprints. Preprints were included to capture the recent developments in LLM and agentic biomedical summarization.

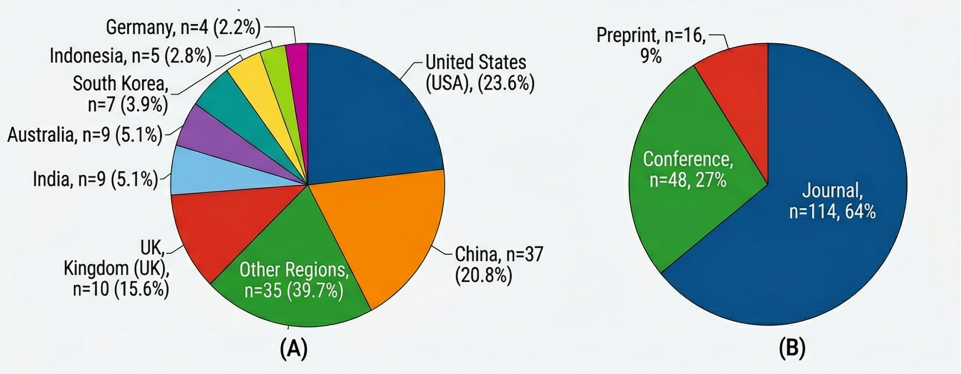

Figure 3: Geographic and publication characteristics of the analyzed corpus (N = 178). (**A**) Global distribution of biomedical summarization research output. (**B**) Distribution of publication venue types.

3.1.2 Dataset and Domain Distribution

Cross-tabulation revealed a significant association between clinical subdomains and data provenance (Table 6). The audited corpus shows clusters of data sources, clinical domains, and reporting practices that may affect the generalizability and reproducibility of current biomedical summarization research. Radiology emerged as the most frequently represented clinical subdomain (66.9%, n = 119) and exhibited the highest utilization of the MIMIC ecosystem, with 47.9% (57/119) of radiology-focused studies employing MIMIC-CXR, MIMIC-III/IV, or related subsets. Other subdomains with more complex narratives, such as Clinical Records & EHR (16.3%) and Pathology (2.8%), used mostly a private institutional dataset. Approximately 34.3% (n = 61) of all included studies utilized the MIMIC Datasets, whereas scientific literature summarization was mostly supported by public resources like PubMed. This institutional reliance introduces a significant structural boundary, which constrains generalizability to other healthcare systems, patient populations, and non-English clinical environments.

Table 6: Cross-tabulation of clinical subdomains and data provenance (N = 178).

Clinical Sub-Domain	Prev. (%)	Total (n)	Public Bench.	MIMIC Family	Private Inst.	Other	Narrative Complexity
Radiology	66.90%	119	51	57	11	0	High
Clinical Records & EHR	16.30%	29	6	4	14	5	Very High
Specialized/Other	5.60%	10	4	0	4	2	Mod. to High
Scientific Literature	3.90%	7	7	0	0	0	High
Pathology	2.80%	5	2	0	3	0	Very High
Ophthalmology	2.20%	4	2	0	2	0	High
Cardiology	2.20%	4	1	2	0	1	High

3.1.3 Risk of Bias and Dataset Characteristic Profile

The audited corpus showed significant clusters across data sources, clinical domains, and reporting practices. Current biomedical summarization research may be affected by these factors in terms of generalizability and reproducibility. The 178 included studies were audited to examine the current research landscape using eight predefined risk indicators. As summarized in Table 7, the majority of studies used English-language data (82.6%) and targeted radiology-related tasks (66.9%). Transparency was limited; only 41.6% of studies released source code, and only 6.2% reported the exact prompts used in evaluation. Furthermore, 34.3% of studies were dependent on the MIMIC ecosystem, which further complicates the issue of external validity. To account for these potential biases, all audit indicators were included as covariates in the multivariable analyses. Each of the eight risk-of-bias indicators was incorporated into both multivariable and sensitivity analyses.

Table 7: Summary profile of study-level risk-of-bias indicators and dataset characteristics (N = 178).

Risk Indicator	(n)	%	High-Risk Impact/Interpretation
English-Language Concentration	147	82.6%	Limits global generalizability to non-Western healthcare settings.
Radiology-Only Task Focus	119	66.9%	Overrepresents patterns specific to high-volume image-text pairs.
Prompt Unavailability	167	93.8%	Critical reproducibility barrier; prevents independent auditing of LLM reasoning.
Code Unavailability	104	58.4%	Restricts independent verification of data processing and engineering pipelines.
MIMIC Ecosystem Dependency	61	34.3%	High risk of institutional over-fitting (the "Generalizability Mirage").
Closed-Weight/API Models	48	27%	Risks "behavior drift" due to unannounced vendor-side updates.
Institutional/Private Data	50	28.1%	Prevents external auditing of safety and clinical accuracy claims.
Preprint Status	16	9%	Presence of non-peer-reviewed evidence within the "Evaluation Rebound".

The cumulative Risk-Flag Count (0–8) and coding criteria for each study is reported in Appendix C Tables A6 and A7. Each of the eight risk-of-bias indicators was incorporated into both multivariable and sensitivity analyses. In the multivariable ordinal logistic regression, all indicators were included as covariates to evaluate the EAS–CRL association after adjustment; none showed a significant independent association with CRL (all $p > 0.05$). In the stratified sensitivity analysis, none of the between-stratum differences in the EAS–CRL correlation remained significant after Holm–Bonferroni correction as will be discussed in Section 3.6.7.

3.1.4 Inter-Rater Reliability, Internal Consistency Baseline

The reliability of the framework proposed by MQS–EAS–CRL is validated by conducting an agreement and consistency analysis across the entire corpus (N = 178). All three instruments demonstrated excellent inter-rater agreement, with ICC (2, 1) values of 0.91 (95% CI: 0.88–0.94) for MQS, 0.89 (95% CI: 0.85–0.92) for EAS, and 0.91 (95% CI: 0.85–0.96) for CRL, all exceeding commonly accepted reliability values (Table 8). Consistency analysis also confirmed the stability of the structure, with KR-20 values of 0.84 for the binary MQS indicators and the Cronbach's α value of 0.78 for the six-dimensional EAS structure. Agreement at the item level was high across individual MQS criteria (κ = 0.72–0.91), and CRL staging demonstrated high agreement at the weighted level (κ = 0.92) and item level (κ = 0.94) for the clinical maturity level (CRL ≥ 3). Detailed item-level agreement results are presented in Appendix D Table A8. Overall, these results indicate that the MQS–EAS–CRL framework is a promising, consistent, and reliable approach for evaluating methodological quality, evaluation adequacy, and clinical readiness in biomedical summarization studies.

Table 8: Psychometric evaluation of framework reliability: inter-rater agreement and internal consistency.

Instrument/ Construct	Metric Type	Statistical Test	Value	95% Confidence Interval	Interpretation
Methodological Quality (MQS)	Inter-Rater Stability	ICC2,1	0.91	[0.88, 0.94]	Excellent Absolute Agreement
	Internal Consistency	KR-20	0.84	N/A	Robust Structural Integrity
	Item-Level Accuracy	Cohen's κ	0.7940	[0.72, 0.85]	Substantial to Near-Perfect
Evaluation Adequacy Score (EAS)	Inter-Rater Stability	ICC2,1	0.89	[0.85, 0.92]	Excellent Absolute Agreement
	Internal Consistency	Cronbach's α	0.78	N/A	High Dimensional Convergence
	Dimension Accuracy	ICC2,1	0.86–0.95	[0.80, 0.98]	Excellent Dimension Stability
Clinical Readiness Level (CRL)	Inter-Rater Stability	ICC2,1	0.91	[0.85, 0.96]	Excellent Absolute Agreement
	Ordinal Agreement	Quadratic κw	0.92	[0.87, 0.96]	Near-Perfect Staging Stability
	Binary Staging Assessment (CRL ≥ 3)	Cohen's κ	0.94	[0.89, 0.99]	Near-Perfect Exploratory level

3.2 Methodological Quality and Data Adequacy

The systematic audit of the included studies by Methodological Quality Score (MQS) indicated strong methodological transparency across the corpus. The median MQS was 7 (IQR: 6–8), with 64.0% (n = 114) of studies categorized as High Quality (MQS 7–8), 33.1% (n = 59) as Moderate Quality (MQS 4–6), and 2.8% (n = 5) only as Low Quality (Appendix D Table A9).

The results of item-level analysis in Table 9 showed that the highest level of compliance was with Model Specification (MT1; 99.4%), External Test Set Usage (DQ2; 94.4%), and Baseline Comparison (ER2; 92.7%), indicating that most studies provide sufficient information regarding model architectures, validation datasets, and comparative benchmarking. Dataset Size Reporting (DQ1) was also common (74.7%), reflecting a generally established—although not universal standard for methodological reporting. Public source code availability (RP1) was reported in only 41.6% (n = 74) of studies, revealing a substantial reproducibility gap within the literature.

Table 9: Methodological transparency audit: item-level compliance rates for the 8-item MQS (n = 178).

MQS Item	Domain	Description	Pass Rate (%)	(n)
MT1	Model Transparency	Model architecture fully specified	99.4%	177
MT2	Model Transparency	Hyperparameters and training details reported	84.8%	151
ER1	Evaluation Rigor	Multiple evaluation metrics reported	85.4%	152
ER2	Evaluation Rigor	Comparison against at least one baseline	92.7%	165
DQ1	Data Quality	Dataset size and split explicitly stated	74.7%	133
DQ2	Data Quality	External or held-out test set used	94.4%	168
RP1	Reproducibility	Code publicly available	41.6%	74
RP2	Reproducibility	Data publicly available or access pathway described	79.8%	142

Reproducibility and Transparency

The items in Table 10 summarize the reporting of transparency, reproducibility, and assessment rigor across the 178 included studies. Overall, data provenance was documented, with over 90% of studies reporting a clear data source or access pathway, reflecting reliance on common resources such as MIMIC-CXR and MIMIC-III. In comparison, artifacts related to reproducibility were reported significantly less often. Only 41.6% (n = 74) of studies provided public source code, and 43.8% (n = 78) of studies reported explicit hallucination taxonomies. Inter-rater agreement statistics of human evaluation were reported in 19.1% of studies (n = 34). Prompt transparency was the most pronounced reporting gap: the exact prompts were shared by only 6.2% (n = 11) of studies, whereas 71.9% (n = 128) provided a general description of the prompting strategy.

Table 10: Reporting prevalence of validation and reproducibility indicators (N = 178).

Reporting Indicator	Category/Status	(n)	(%)
Hallucination Taxonomy	Reported (Explicit Categorization)	78	43.8%
LLM Judge	Yes (LLM-as-a-Judge utilized)	23	12.9%
IAA Reported	Yes (Inter-Rater Agreement reported)	34	19.1%
Code Availability	Publicly Available	74	41.6%
	Available Upon Request/Not Specified	104	58.4%
Data Provenance	Public Benchmark	107	60.1%
	Restricted	28	15.7%
	Institutional	22	12.4%
	Mixed/Upon Request	15	8.4%
	Not Reported	6	3.3%
Prompt Engineering	Exact Prompts Shared	11	6.2%
	Strategy Described	128	71.9%

Note: The Hallucination Taxonomy row (n = 78, 43.8%) reports the count of studies that applied an explicit multi-label hallucination taxonomy. This sub-corpus is a strict subset of the broad clinical-failure-reporting group (n = 99) reported in next Section 3.3.1.

This study provides further evidence for the lack of reproducibility in biomedical summarization research. Although dataset provenance is often reported, a few key artifacts, such as source code, prompt configurations, protocols for hallucination auditing, and measures of reliability in human evaluation, are

not sufficiently reported. This lack of transparency introduces issues with replicability and cross-study comparison and impedes accurate evaluation of clinical deployment readiness.

3.3 Evaluation Adequacy Score (EAS) Results

The six-dimensional Evaluation Adequacy Score (EAS) was extracted from the included studies to quantify the depth of evaluation beyond metric adoption. The median EAS of 0.46 (IQR: 0.42–0.58) indicated that most studies covered fewer than half of the critical evaluation dimensions. The distribution of EAS dimensions reveals a marked imbalance between technical evaluation and clinically grounded validation. Clinical Validity (D5) was the most frequently reported dimension (89.3%, n = 159), followed by Safety Auditing (D6; 88.2%, n = 157) and Factuality & Grounding (D3; 77.0%, n = 137) as shown in Table 11. Semantic Alignment (D2) was reported in only 38.2% (n = 68) of studies, and rigorous assessment in only 2.8% (n = 5).

Table 11: Multi-dimensional distribution of evaluation adequacy across scoring tiers (N = 178).

EAS Dimension Axis	Score 0 (Absent)	Score 1 (Partial)	Score 2 (Rigorous)	Total Adoption % (n)
D1: Lexical Metrics	46	24	108	74.2% (132)
D2: Semantic Alignment	110	63	5	38.2% (68)
D3: Factuality & Grounding	41	114	23	77.0% (137)
D4: Human Expert Eval	78	60	40	56.2% (100)
D5: Clinical Validity	19	102	57	89.3% (159)
D6: Safety Audit	21	125	32	88.2% (157)

Lexical Metrics (D1) were the most mature technical evaluation component, 74.2% (n = 132) of studies used multiple benchmark measures. Human Expert Evaluation (D4) was reported in 56.2% (n = 100) of studies, while 43.8% (n = 78) reported no formal expert evaluation. Although safety considerations were frequently discussed, relatively few studies conducted structured, quantitative safety audits, highlighting a persistent gap between reported performance and clinically meaningful evaluation.

Overall, the EAS distribution highlights a persistent evaluation-depth gap within the biomedical summarization literature. While lexical performance, factuality, clinical validity, and safety are increasingly reported, rigorous semantic evaluation, structured expert assessment, and comprehensive safety auditing remain comparatively uncommon, limiting confidence in the clinical readiness of many proposed systems.

3.3.1 Safety and Bias Audit Deficit

A systemic Safety Gap persists across the total corpus (N = 178). Although 88.2% (n = 157) of studies performed at least a partial safety audit (EAS D6 ≥ 1), only 18.0% (n = 32) conducted rigorous safety audits that included structured hallucination analysis, risk assessment, or formal error evaluation. These figures contrast with the 84.3% (n = 150) of studies that qualitatively discussed safety concerns within their narrative analysis, a thematic prevalence further investigated in Section 3.4.1. To characterize these risks, a multi-label analysis was performed on two distinct levels of the corpus based on the depth of reporting:

- General Clinical Failure Reporting (n = 99): Among studies that reported any clinical failure mode or hallucination (either qualitatively or quantitatively), Entity Fabrication was the most frequent error (45.5%), followed by Omission (38.4%) and Relation Distortion (34.3%), as detailed in Table 12.

- Explicit Taxonomic Auditing (n = 78): Within the specialized sub-corpus of studies utilizing explicit, multi-labeled hallucination taxonomies, the detection sensitivity for errors was significantly higher. In this group, Omission (66.7%) and Entity Fabrication (61.5%) were identified as the most pervasive failure modes.

Table 12: Frequency distribution of reported clinical failure modes and hallucination categories (n = 99).

Rank	Error Category	n	(%)	Operational Definition
1	Entity Fabrication (EF)	45	45.5%	Invention or incorrect generation of clinical entities.
2	Omission (OS)	38	38.4%	Failure to include clinically relevant information present in the source document, potentially affecting summary completeness.
3	Relation Distortion (RD)	34	34.3%	Distortion of relationships between clinical entities, including negation reversals, incorrect temporal relations, or semantic inconsistencies.
4	Factual Inaccuracy (FI)	22	22.2%	Statements that contradict source information or misrepresent clinical facts, including anatomical or diagnostic misinterpretation.
5	Temporal Error (TE)	3	3.0%	Failures in temporal sequencing, misattribution of symptom onset, or multi-document integration errors.
—	Total Tracked Multi-Labels	142	143.4%	Reflects the total number of category tags applied across the 99 analyzed studies.

Note: The n = 99 denominator represents the broad reporting group—studies that discussed any clinical failure or hallucination. For detection frequencies within the high-sensitivity sub-group utilizing structured taxonomies (n = 78), see previous Section 3.3.1 (Table 10).

This divergence suggests that while general reporting identifies broad trends, structured auditing pipelines are characteristically associated with increased sensitivity to complex clinical errors. As summarized in Table 12, multiple error categories frequently occurred within the same study, resulting in a cumulative tagging rate of 143.4% across the 99 reported studies.

3.3.2 EAS Dimension Level Associations with Clinical Readiness Level (CRL)

Table 13 presents the dimension-level associations between the six Evaluation Adequacy Score (EAS) components and clinical readiness Level (CRL). The dimension-level analysis shows a significant alignment between specific evaluation components and Clinical Readiness. Safety and Bias Auditing indicated the highest correlation with translational maturity ($\rho = 0.76$), followed by Clinical Utility ($\rho = 0.681$). These correlations are descriptive of the corpus and partly reflect the overlap between the EAS dimensions (D5, D6) and the criteria used to score CRL stages. Factuality evaluation showed a moderate correlation ($\rho = 0.544$) and Human Expert Assessment a significant but smaller one ($\rho = 0.421$).

Most significantly, Lexical Metrics had the lowest correlation with Clinical Readiness (0.182, $p = 0.115$), and the 95% confidence interval for Lexical Metrics crossed zero (−0.024 to 0.353), indicating that the association is not statistically reliable. These findings indicate that strong linguistic similarity to reference summaries does not necessarily translate into clinically trustworthy or deployment-ready systems. Many studies demonstrated acceptable results on ROUGE- and BLEU-based measures, with little evidence of safety

auditing, factual verifications, clinical validation, or formal expert evaluation. Fig. 4 illustrates that strong lexical performance was often accompanied by clinically relevant deficits such as hallucinations, omissions, and poor safety evaluation.

Table 13: Dimension-level correlations with clinical readiness level (CRL).

Evaluation Dimension	Spearman ρ	Exact *p*-Value	95% CI [Boot-Strap]	Interpretation
Safety & Bias (Saf)	0.762	<0.0001	[0.684, 0.821]	Strongest Association: Reflects high procedural alignment with the safety requirements of mature clinical systems.
Clinical Utility (Cli)	0.681	<0.0001	[0.592, 0.748]	Strong Association: Strongly linked to high CRL (Levels 4–5).
Factuality (Fac)	0.544	<0.0001	[0.435, 0.632]	Moderate Association: Essential baseline for clinical safety.
Human Expert (Hum)	0.421	0.0014	[0.306, 0.518]	Moderate Association: Significant but secondary to safety auditing.
Semantic Similarity (Sem)	0.283	0.0245	[0.118, 0.407]	Weak Association: Often inconsistent across different medical tasks.
Lexical Metrics (Lex)	0.182	0.1154	[−0.024, 0.353]	Negligible: no significant association for safety.

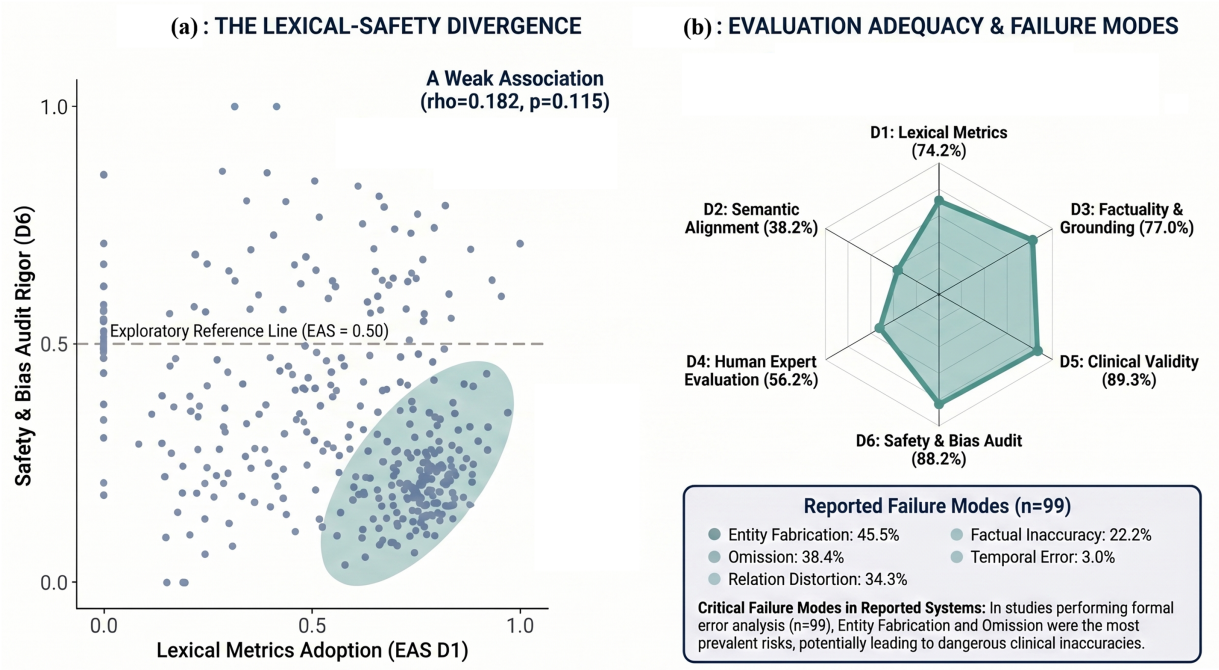

Figure 4: Visualization of the lexical-safety divergence N = 178.

The lexical safety divergence is apparent in Fig. 4A, where many studies had high lexical evaluation scores but little coverage of the safety audit. This shows a persistent clinical safety gap. Studies closer to higher stages of clinical readiness, by contrast, consistently emphasized safety auditing and clinically grounded validation. Furthermore, Receiver Operating Characteristic (ROC) modeling identified an internally derived exploratory reference point at an EAS value of 0.50 (AUC = 0.81, 95% [CI: 0.42, 0.58]). Operating at this mathematically exploratory reference point yields a high clinical exclusion Specificity of 0.79, successfully identifying systems characteristically associated with institutional or prospective pilots. Fig. 4B also indicates that while Clinical Validity, Safety Auditing, and Factuality Assessment are common, there are significant gaps in evaluation depth, especially in factuality assessment and structured human expert Assessment.

3.4 Clinical Readiness and Deployment Constraints

To assess progression from experimental development to clinical integration, studies were classified using the five-level Clinical Readiness Level (CRL) framework (Table 14). The resulting distribution shows a clear translational funnel, with the majority of studies (75.3%, n = 134) at CRL 1–2, relying on either benchmark-based evaluation or limited expert review. Retrospective institutional validation was used to reach CRL 3 by 19.1% (n = 34), and prospective clinical pilots were used to reach CRL 4 by 5.1% (n = 9). Remarkably, one study (0.6%) received a CRL of 5, indicating the implementation of an EHR-integrated summarization system in routine clinical practice [92]. To ensure statistical stability and mitigate the impact of sparse high-level samples (CRL 4–5, n = 10), the primary inferential focus is placed on a binary stratification: early-stage laboratory systems (CRL 1–2, n = 134) vs. translationally advanced systems (CRL ≥ 3, n = 44). Results involving CRL 4 and CRL 5 are treated hereafter as descriptive supplementary evidence of emerging implementation pathways rather than as a basis for robust statistical inference.

Table 14: Distribution of translational maturity across the five clinical readiness levels (N = 178).

CRL Level	Designation	(%)	Primary Characteristics
CRL 1	Benchmark	23.6%	Public benchmark evaluation with automated metrics only.
CRL 2	Expert Review	51.7%	Public or retrospective data with informal or limited expert review.
CRL 3	Institutional	19.1%	Institutional or private clinical data with structured expert evaluation.
CRL 4	Prospective	5.1%	Prospective pilot in a real or simulated clinical workflow.
CRL 5	Operational	0.6%	Sustained routine clinical deployment.

The results show a substantial disconnect between innovation and clinical implementation, despite ongoing progress in Transformer-, LLM-, and agentic-based summarization systems. The clinical readiness funnel illustrated in Fig. 5 was consistently observed across architectural families and technological eras, suggesting that barriers to clinical deployment remain common structural characteristics of the current biomedical summarization literature.

3.4.1 Safety, Bias, and Regulatory Compliance

An analysis of the three dimensions of safety, fairness, and regulatory governance was conducted because they are considered essential elements in clinical adoption of biomedical summarization systems. The number of studies that addressed safety concerns was great (84.3%, n = 150), but only a small fraction (18%, n = 32) carried out thorough and quantified safety audits that included structured analysis of hallucinations, categorization of errors, or formal risk assessment (Table 15). Similarly, very few studies (42.7%, n = 76) considered potential concerns of demographic bias. Regulatory and governance issues were

reported even less frequently (33.1%, n = 59), with only a small number explicitly mentioning Institutional Review Board (IRB) approval, Health Insurance Portability and Accountability Act (HIPAA), or General Data Protection Regulation (GDPR) compliance.

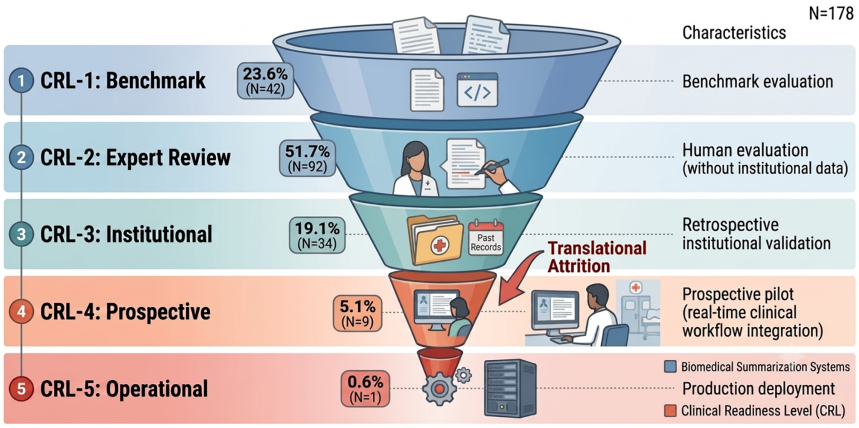

Figure 5: Clinical readiness funnel of the included biomedical summarization studies across CRL 1–5.

Table 15: Depth of safety and governance evaluation (N = 178).

Reporting Domain	Explicit Assessment (n, %)	Not Addressed/NR (n, %)
Safety Analysis (Qualitative)	150 (84.3%)	28 (15.7%)
Quantified Safety Audit	32(18%)	146 (82.0%)
Demographic Bias Assessment	76 (42.7%)	102 (57.3%)
Regulatory Compliance (HIPAA/GDPR/IRB)	59 (33.1%)	119 (66.9%)

This shows that the current evaluation process still emphasizes technical performance over the broader safety and governance considerations needed for real-world clinical use. This imbalance suggests that evaluation practices continue to emphasize technical performance over the broader safety, ethical, and governance requirements necessary for clinical deployment, reinforcing the translational barriers identified in the Clinical Readiness analysis.

3.4.2 Thematic Analysis of Deployment Barriers: Technical vs. Operational

To identify the main factors limiting the translation of biomedical summarization systems into clinical practice, implementation barriers reported across the corpus were grouped into four thematic domains (Table 16). Technical barriers were the most reported challenge, appearing in 65.2% of studies (n = 116), including hallucinations, reliability limitations, computational cost, explainability deficits, context-window constraints, and multimodal integration challenges. Operational barriers (12.9%, n = 23) mainly centered on workflow integration, EHR interoperability, clinician acceptance, implementation challenges, and data access issues. Ethical and regulatory concerns were reported in 14.0% (n = 25) of studies and covered privacy, fairness, transparency, accountability, and compliance.

This highlights the critical need not only to enhance model performance but also to develop integration, governance, and implementation requirements.

Table 16: Primary implementation barriers by thematic domain (N = 178).

Barrier Domain	% (n)	Representative Barrier Types
Technical	65.2% (116)	Hallucinations, model reliability, computational cost, explainability, context limitations, multimodal integration
Accuracy/Clinical Validity	7.9% (14)	Factual errors, omissions, diagnostic inaccuracies, summary consistency, clinical correctness
Operational	12.9% (23)	Workflow integration, EHR interoperability, clinician acceptance, deployment logistics, data availability
Ethical/Regulatory	14.0% (25)	Privacy, fairness, bias, transparency, accountability, HIPAA/GDPR compliance, governance requirements

3.4.3 EAS–CRL Correlation Analysis

A Spearman rank correlation was calculated between the Evaluation Adequacy Score (EAS) and Clinical Readiness Level (CRL) to assess if evaluation depth is correlated with translational readiness, for the N = 178 studies. A bias-corrected and accelerated (BCa) bootstrap resampling with 10,000 iterations was used to create the 95% confidence interval. There was a significant positive correlation between evaluation rigor and clinical maturity ($\rho = 0.54$, $p < 0.001$, 95% bootstrapped CI: [0.44, 0.63]), reflecting the procedural alignment rather than an independent predictive relationship between the EAS and CRL constructs (Fig. 6). This correlation ($\rho = 0.54$) highlights that evaluation depth and translational staging are conceptually interdependent constructs.

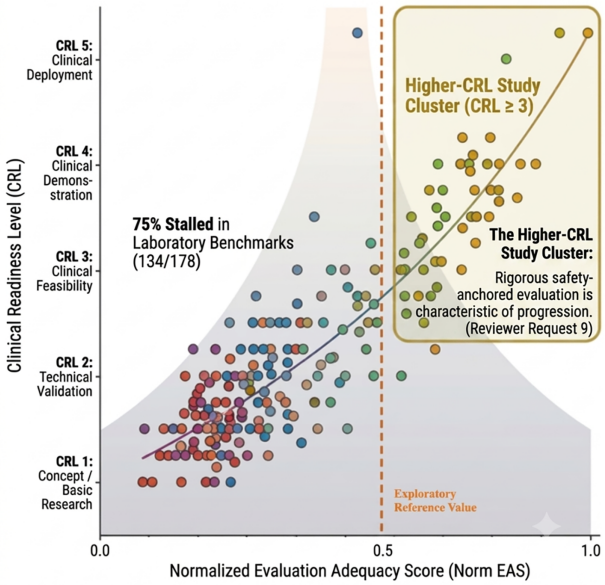

Figure 6: Scatter plot of evaluation adequacy score (EAS) vs. clinical readiness level (CRL) (N = 178).

A post hoc analysis revealed that studies with an EAS below 0.33 were mostly in CRL 1–2, and most studies reaching institutional feasibility (CRL ≥ 3) did so with an EAS at or above 0.50. Within the analyzed corpus, studies that reached CRL ≥ 3 tended to cover at least half of the EAS dimensions with substantive

depth; this pattern is descriptive of the present sample and does not constitute a deployment requirement. This association was more pronounced in hybrid/agentic systems (ρ = 0.78) than in decoder-only systems (ρ = 0.42), which may be attributed to greater validation needs and the built-in self-correction inherent in multi-agent pipelines. In higher quality studies (MQS 7–8), the discriminative power of the EAS was strong (AUC = 0.835), and the previously established exploratory reference point was held: while studies reaching institutional feasibility (CRL ≥ 3) characteristically demonstrated an advanced validation depth (EAS ≥ 0.50). Thus, while studies reaching institutional feasibility (CRL ≥ 3) characteristically demonstrated an advanced validation depth (EAS ≥ 0.50) with only few exceptions studies (*n* = 6), notably none of the studies with an EAS below the exploratory reference point exceeded CRL 2. This density pattern is shown in Fig. 7 as a heatmap with the exploratory referenced point of readiness set at EAS = 0.50.

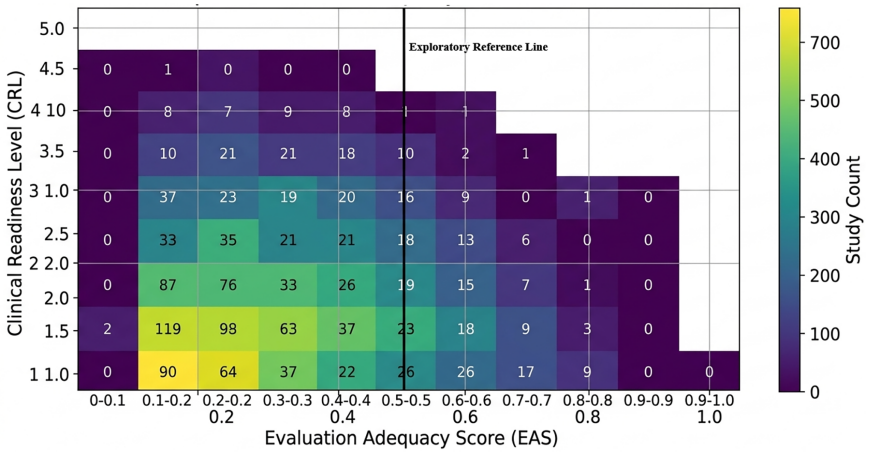

Figure 7: Heatmap of evaluation adequacy score (EAS) vs. clinical readiness level (CRL) (N = 178).

Two outlier patterns warrant comment. Study [93] achieved the most comprehensive evaluation in the corpus (normalized EAS: 1), yet remained at the technical validation stage CRL-2. This demonstrates that while a multidimensional evaluation is essential, translational progress is often decoupled from technical rigor by regulatory and operational barriers. Conversely, a small group of studies attained institutional feasibility (CRL-3) despite reporting qualitative pilot designs with lower evaluation depth (normalized EAS: 0.25–0.33) such as studies [94–96]. These results are consistent with a procedural alignment between evaluation methodology and translation maturity within the analyzed corpus. The corpus-level pattern observed in hybrid/agentic studies, in which higher evaluation adequacy co-occurred with higher CRL stages, is descriptive and may reflect partial construct overlap between EAS dimensions D4–D6 and the criteria used at higher CRL stages, rather than an inherent superiority of any architectural family.

3.5 Technical Architecture and Performance Landscape

Due to considerable heterogeneity across studies, the following comparisons reflect general trends rather than definitive comparative effectiveness. The summarization models across the corpus were classified using a two-tier scheme Backbone Architecture and System/Pipeline Design, as summarized in Table 17. A clear transition is taking place from earlier encoder-decoder approaches toward large-scale generative models and increasingly complex workflow designs. At the architecture level, the most frequently used backbone architecture was Decoder-only (50.6%, n = 90), followed by Hybrid/Agentic (26.4%, n = 47), Encoder–Decoder (19.7%, n = 35), and Encoder-only (3.4%, n = 6). GPT-4, Llama-3, Mistral, BioMistral, BART, T5, PEGASUS, and BioBART were the most representative models. At the system-design level, the

dominant approach was the Abstractive summarization (75.3%, n = 134), followed by Hybrid pipelines integrating RAG or agentic workflows (21.9%, n = 39), with Extractive approaches accounting for the remaining 2.8% (n = 5).

Table 17: Integrated analysis of architectural taxonomy, system design, and translational outcomes (N = 178).

Taxonomy Level	(n)	%	Mean EAS	Median EAS	Mean CRL	CRL ≥ 3 n (%)
Tier 1: Backbone Architecture						
Decoder-only	90	50.60%	0.499	0.498	2.056	27 (30.0%)
Hybrid/Agentic	47	26.40%	0.46	0.459	2.043	8 (17.0%)
Encoder–Decoder	35	19.70%	0.459	0.459	2.171	9 (25.7%)
Encoder-only	6	3.40%	0.433	0.433	2	0 (0.0%)
Tier 2: System Design						
Abstractive	134	75.30%	0.476	0.466	2.075	36 (26.9%)
Hybrid (Agentic/RAG)	39	21.90%	0.507	0.533	2.077	7 (17.9%)
Extractive	5	2.80%	0.318	0.368	2	1 (20.0%)
Overall Corpus	178	100%	0.478	0.46	2.073	44 (24.7%)

Among decoder-only models (n = 90), the most common optimization approach was supervised adaptation, including parameter-efficient methods such as PEFT and LoRA (50.0%, n = 45). Prompt-based methods accounted for 40.0% of the corpus, comprising prompt-only methods (17.8%), zero-shot (14.4%), and few-shot (5.6%). By contrast, Chain of Thought (CoT) prompting was reported in only 2.2% of studies (n = 2), while instruction tuning was reported in only 1.1% (n = 1) (Appendix D Table A10).

Furthermore, Table 17 integrates the architectural taxonomy and system design with the corresponding evaluation adequacy and clinical readiness values observed within the analyzed corpus. Because there is substantial variation in summarization tasks, clinical domains, datasets, and evaluation methodologies across studies, the results below are descriptive and should not be interpreted as estimates of summarization effectiveness or direct comparisons of model architectures.

At the backbone level, the observed median EAS values were 0.498 for decoder-only studies (n = 90), 0.459 for hybrid/agentic studies (n = 47), 0.459 for encoder–decoder studies (n = 35), and 0.433 for encoder-only studies (n = 6). The percentage of studies placed in the CRL ≥ 3 category was 30.0% (27/90) for decoder-only, 17.0% (8/47) for hybrid/agentic, 25.7% (9/35) for encoder–decoder, and 0% (0/6) for encoder-only. There are 6 studies in the encoder-only stratum, so the CRL ≥ 3 count is actually observed and not evidence of a between-family difference.

At the system-design level, abstractive studies (n = 134) reported a median EAS of 0.466 and a mean CRL of 2.075, with 26.9% (36/134) classified at CRL ≥ 3. Hybrid (agentic/RAG) studies (n = 39) reported a median EAS of 0.533 and a mean CRL of 2.077, with 17.9% (7/39) at CRL ≥ 3. Extractive studies (n = 5) reported a median EAS of 0.368 and a mean CRL of 2.0, with 20.0% (1/5) at CRL ≥ 3.

The values in the extractive model are small (n = 5) and should be interpreted as descriptive observations rather than as stable estimates. The patterns are reported descriptively. The above values should not be interpreted as an order of effectiveness across architectures or across studies within each architecture because ROUGE-L and BERTScore are not interpretively equivalent across radiology, EHR notes, and scientific-literature summarization tasks, and evaluation protocols vary between studies in each architectural stratum.

3.5.1 Performance Benchmarks by Architecture Family and Text Type

Table 18 summarizes reported lexical overlap (ROUGE-L) and semantic similarity (BERTScore) values, graded by architecture family and clinical text type. Among studies with available metrics, hybrid/agentic systems showed higher median values than encoder-only systems across selected text categories. However, these patterns should be interpreted descriptively because of the substantial diversity across the 178 studies regarding their specific summarization tasks, datasets, and evaluation protocols, and the automated scores (ROUGE-L and BERTScore) are without equivalent meanings across different clinical domains, such as radiology reports (Mdn ROUGE-L = 0.385; Mdn BERTScore = 0.882), EHR notes (Mdn BERTScore = 0.901), pathology reports, and scientific literature. Reported median scores for decoder-only configurations varied widely across text categories, Mdn ROUGE-L = 0.313, and Mdn BERTScore = 0.735 in Radiology Reports, and BERTScore variance was particularly large for general biomedical datasets. Because ROUGE-L and BERTScore have non-equivalent interpretations across radiology reports, EHR notes, pathology reports, and scientific literature summarization, these values are reported descriptively, and direct cross-task or cross-architecture ranking is not warranted.

Table 18: Descriptive meta-summary of reported lexical (ROUGE-L) and semantic (BERTScore) metrics.

Architecture Family	Radiology Reports	Clinical Notes/EHR	General Biomedical	Pathology Reports
Encoder-only	R: 0.191 [0.18–0.21] B: NR	R: NR B: NR	R: 0.221 [0.22–0.22] B: NR	NR
Encoder–Decoder	R: 0.395 [0.36–0.41] B: 0.747 [0.75–0.75]	R: 0.518 [0.51–0.52] B: 0.552 [0.55–0.55]	R: 0.538 [0.39–0.68] B: NR	NR
Decoder-only	R: 0.313 [0.27–0.41] B: 0.735 [0.51–0.85]	R: 0.453 [0.43–0.47] B: NR	R: 0.246 [0.20–0.29] B: 0.857 [0.86–0.86]	R: 0.579 [0.58–0.58] B: NR
Hybrid/Agentic	R: 0.385 [0.30–0.41] B: 0.882 [0.82–0.89]	R: 0.492 [0.31–0.60] B: 0.901 [0.90–0.91]	R: 0.339 [0.34–0.34] B: 0.774 [0.77–0.77]	NR

Note: R = ROUGE-L; B = BERTScore. NR = Not Reported. All values standardized to a (0.0–1.0) scale, Values calculated as Median [25th–75th percentile].

In an encoder–decoder architecture, dense sequence-to-sequence modeling is advantageous for mimicking medical terms and sentence structure. In Radiology Reports, they obtain a peak median ROUGE-L of 0.395 [0.36–0.41], and their absolute maximum is in General Biomedical literature (Mdn ROUGE-L = 0.538). Their performance on dense scientific corpora, such as PEGASUS on PubMed, is indeed high quality, but this is not reflected in their performance in actual medical record workflows (Mdn BERTScore = 0.552). The ROUGE-L scores of the encoder-only models, which are constrained by structural limitations, are 0.191 [0.18–0.21] for Radiology and 0.221 [0.22–0.22] for General Biomedical, respectively. Because ROUGE-L and BERTScore have non-equivalent interpretations across radiology reports, EHR notes, pathology reports, and scientific literature summarization, these values are reported descriptively, and direct cross-task or cross-architecture ranking is not warranted.

Descriptive inferential tests were used to describe variation within the corpus, but not to establish a comparative ranking. The EAS distribution was subjected to a one-way Analysis of Variance (ANOVA) across the architectural paradigms that returned $F(2, 175) = 6.47$, $p = 0.0019$, and the CRL distribution was subjected to a Kruskal–Wallis test that returned $H = 7.12$, $p = 0.028$. These results indicate that both EAS and CRL distributions varied significantly across the architectural strata presented in Table 17. Based on studies that reported metrics, the median EAS in the hybrid/agentic system was 0.533, whereas the abstractive and extractive strata had median EAS values of 0.466 and 0.368, respectively. The statistical

tests above are reported as descriptive evidence of within-corpus heterogeneity rather than as comparative-effectiveness estimates.

3.5.2 Longitudinal Trends and Architectural Safety Benchmarks

To evaluate temporal evolution in evaluation rigor and clinical translation, a Kruskal–Wallis H test was applied across the three publication eras: Era 1 (2017–2020; n = 2), Era 2 (2021–2023; n = 41), and Era 3 (2024–2026; n = 135). Post-hoc analyses showed a significant temporal increase in evaluation rigor (EAS: H = 17.2145, p = 0.0002) and clinical readiness level (CRL: H = 11.8942, p = 0.0026). The results of the post-hoc analysis showed that the EAS of the current generative and agentic systems (Era 3, median = 0.50) is significantly higher than that of Era 2 (0.42, p = 0.0084) and Era 1 (0.42), revealing a higher median EAS in Era 3 studies than in earlier eras within this corpus (Appendix D, Table A11).

A Chi-square test further showed a significant association between architecture type and safety auditing practices ($\chi^2(2) = 13.5182$, $p = 0.0012$, V = 0.2756). The percentage of studies that conducted a structured safety audit was 91.5% in hybrid/agentic studies, 67.8% in decoder-only studies, and 51.4% in encoder–decoder studies (Table 19). These ratios are representative of reporting in the studies in the corpus. They do not, on their own, distinguish between two non-exclusive explanations: (i) that there is a real difference in the practice of safety audits across different architectural strata, and (ii) that the reporting convention effect occurs, in which the inclusion of structured safety-audit sections is more common in more recent agentic and RAG-based publications. Both explanations may be considered consistent with the observed values; the present corpus lacks the design elements needed to distinguish between them.

Table 19: Architectural safety audit rigor: 3 × 2 contingency table (N = 172, excluding the 6 encoder-only).

Architecture Family	No Safety Audit (D6 = 0)	Safety Audit (D6 ≥ 1)	Total Cohort (n)	Definitive Audit Rate (%)
Encoder–Decoder	17	18	35	51.40%
Decoder-only	29	61	90	67.80%
Hybrid/Agentic	4	43	47	91.50%
Aggregate Corpus	50	122	172	70.90%

3.6 Framework Validation and Integrative Analysis

To evaluate the robustness and practical utility of the proposed MQS–EAS–CRL measurement infrastructure, a multi-layer validation protocol was conducted. Validation encompassed conceptual alignment with established clinical AI standards, psychometric reliability assessment, construct and criterion validity testing, modeling, calibration, and robustness analyses. Together, these procedures were designed to determine whether methodological quality (MQS), evaluation adequacy score (EAS), and clinical readiness level (CRL) constitute distinct yet complementary constructs that explain translational progression in biomedical summarization research.

3.6.1 Conceptual Alignment with Clinical AI Standards

The validity of the proposed framework was first assessed by aligning it with internationally recognized standards for trustworthy clinical AI, including CLAIM, CONSORT-AI, Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence (SPIRIT-AI), TRIPOD-AI, DECIDE-AI,

and World Health Organization (WHO) governance recommendations (Table 4). The mapping demonstrated substantial conceptual correspondence across the clinical AI lifecycle. MQS captures methodological transparency and reporting rigor; EAS operationalizes multidimensional evaluation quality across lexical, semantic, factual, clinical, and safety dimensions; and CRL measures translational maturity from laboratory experimentation to routine deployment. Collectively, the three components provide complementary coverage of development quality, evaluation rigor, and clinical maturity.

3.6.2 Empirical Validation through Framework Convergence

Additional support for the framework's validity emerged from convergence across independent analyses. Studies with stronger methodological reporting generally demonstrated more comprehensive evaluation practices, while higher evaluation adequacy scores were consistently associated with greater clinical readiness. Importantly, clinically grounded dimensions, including clinical validity, expert evaluation, factual consistency, and safety auditing, showed substantially stronger relationships with readiness than traditional lexical performance metrics. Similar patterns were observed across architectural analyses, deployment barrier assessments, and readiness profiling, providing convergent descriptive evidence that, within the analyzed corpus, translational maturity co-varied more closely with evaluation depth than with benchmark performance; causal interpretation is not warranted.

3.6.3 Reliability and Internal Consistency

The MQS–EAS–CRL framework is well-structured and internally consistent throughout the corpus, with high reproducibility at each level of measurement (as shown in Tables 8 and A7). The overall inter-rater reliability was adequate, with ICC (2,1) values of 0.91 for MQS and 0.89 for EAS. Cohen's kappa for MQS binary indicators at the item level ranged from 0.72 to 0.91 (0.79). Overall, there was near perfect agreement for Dataset Size (DQ1: $\kappa = 0.91$, 95% CI: [0.85–0.97]), Model Specification (MT1: $\kappa = 0.89$), Source Code Access (RP1: $\kappa = 0.88$), and Raw Data Access (RP2: $\kappa = 0.86$); slightly lower, but still substantial, agreement for Competitive Baselines (ER2: $\kappa = 0.75$, 95% CI: [0.68–0.82]) and Multi-Metric Evaluation (ER1: $\kappa = 0.72$) due to increased subjective judgment in comparative evaluation.

Inter-rater reliability was very good for the continuous EAS dimensions (ICC 0.86 to 0.95). Highest agreement was found in Safety Audit Adherence (ICC = 0.95, 95% CI: [0.91–0.98]) and Lexical Metrics (ICC = 0.94), followed by Semantic Alignment (ICC = 0.92), Human Expert Evaluation (ICC = 0.90), Factuality & Grounding (ICC = 0.88) and Clinical Validity (ICC = 0.86, 95% CI: [0.80–0.91]), demonstrating robust scoring across clinically complex dimensions.

Internal consistency also supports the framework's strength; for MQS, the KR-20 coefficient was 0.84. The multi-axis EAS matrix demonstrated robust scale reliability with a unified Cronbach's $\alpha = 0.7842$. Near-perfect ordinal agreement (weighted $\kappa = 0.92$) was attained by the CRL staging system, especially at the clinically critical level of CRL ≥ 3 ($\kappa = 0.94$). Taken together, these results demonstrate that MQS–EAS–CRL is a reliable, reproducible, and structurally coherent evaluation framework for biomedical summarization systems.

3.6.4 Construct and Criterion Validity

Construct validity analysis demonstrated strong and statistically significant relationships among methodological quality (MQS), evaluation adequacy score (EAS), and clinical readiness level (CRL). All pairwise associations were positive and significant ($p < 0.001$), supporting the conceptual structure of the proposed framework. The strongest relationship was observed between EAS and CRL ($\rho = 0.54$, 95% CI:

[0.44, 0.63]), indicating that evaluation adequacy is the construct most closely aligned with translational maturity. MQS showed a moderate association with CRL (ρ = 0.448, 95% CI: [0.34, 0.54]), and a strong relationship with EAS (ρ = 0.489, 95% CI: [0.38, 0.58]) (Table A12). The high correlation between EAS and CRL (ρ = 0.54) is partially attributable to shared criteria regarding human expert involvement and safety auditing. Consequently, these results should be interpreted as evidence of convergent validity within a unified audit pipeline rather than as evidence of a relationship between fully independent variables.

Standard validity was further demonstrated through significant discrimination between clinical maturity tiers. Systems classified as clinically mature (Tier B; CRL 3–5, n = 44) achieved significantly higher MQS and EAS scores than early-stage systems (Tier A; CRL 1–2, n = 134) ($p < 0.0001$ for both comparisons) (Table A13). Advanced configurations showed significantly higher reporting quality (Tier B Median MQS = 7.0 vs. Tier A Median MQS = 5.0; U = 1820.5, Z = −5.12, $p < 0.0001$) and validation depth (Tier B Median EAS = 0.58 vs. Tier A Median EAS = 0.42; U = 1344.0, Z = −6.84, $p < 0.0001$) (Table A13). The larger effect size observed for EAS (r = 0.51) relative to MQS (r = 0.38) indicates that evaluation adequacy is a stronger discriminator of clinical maturity than reporting quality alone. While the multivariable model was specified across CRL Levels 1–5, the interpretation of factors associated with maturity is targeted toward the transition to institutional feasibility (CRL ≥ 3). This approach ensures that the large effect size (OR = 123.97) is not overinterpreted based on the limited sample size of pilot and production deployments (CRL 4–5).

3.6.5 Associations Validation and Confounder Control

An ordinal logistic regression model was developed to identify an independent relationship of clinical readiness while controlling for methodological quality, publication year, architecture family, and dataset provenance (Appendix E, Table A14).

Multivariable modeling shows that, among the variables examined, EAS exhibits the greatest covariation with clinical readiness (Odds Ratio (OR) = 123.97, $p < 0.0001$); given the shared scoring logic of D4–D6 with higher CRL stages, this estimate should be interpreted as an alignment effect rather than an independent predictive coefficient. This relationship remained robust after adjusting for all confounders, including Publication Year ($p = 0.5076$) and Dataset Provenance ($p = 0.5164$), neither of which had a statistically significant effect (Table A13). Methodological Quality (MQS) was also significantly associated with clinical readiness (OR = 1.51, $p = 0.0008$), as was the use of Hybrid/Agentic architectures (OR = 2.36, $p = 0.0410$). In an auxiliary model, Clinical Team Involvement showed a significant positive contribution to translational maturity (OR = 3.38, $p = 0.004$). These findings are further supported by partial correlation analysis, confirming a strong, confounder-independent association between EAS and CRL (ρ = 0.5212, $p < 0.0001$).

Overall, evaluation adequacy showed the most consistent corpus-level alignment with clinical readiness among the examined descriptors; the alignment was contributed by temporal, architectural, and dataset-related variables. Because the EAS and CRL instruments share criteria for expert validation and safety auditing, this finding is interpreted as a descriptive association within the analyzed corpus, not as an independent predictive determinant.

3.6.6 Validation Value for Clinical Readiness

ROC-based exploratory reference analysis revealed that EAS values were descriptively aligned with CRL ≥ 3 status in the analyzed corpus (AUC = 0.814; 95% CI, 0.742–0.887). A normalized EAS of 0.50 (6/12) emerged as a corpus-derived exploratory reference value at the maximum of the Youden Index, with bootstrap resampling (10,000 iterations) confirming its internal stability (95% CI 0.417–0.583) (Table A15). At this reference value, sensitivity and specificity for CRL ≥ 3 classification within the present corpus were 70.9%

and 78.9%, respectively; these figures characterize internal corpus behavior and should not be interpreted as a generalizable screening threshold (Table A16).

Multivariable ordinal logistic regression across the included studies showed that EAS values co-varied with CRL within the present corpus. The large effect size reported with (OR = 123.97, $p < 0.0001$) reflects the partial overlap in scoring logic between EAS dimensions D4–D6 and the criteria used for scoring CRL levels, while methodological quality (MQS) remains a significant contributor (OR = 1.51, $p = 0.0008$), even after controlling for evaluation rigor, architectural complexity, and publication era (Table A17) Together, these findings support the (EAS ≥ 0.50) as a corpus-derived exploratory reference point descriptively associated with translational maturity in this study.

3.6.7 Framework Robustness and Sensitivity Analyses

This section evaluates the stability of the MQS–EAS–CRL framework under variations in weighting schemes, reporting quality, dataset environments, calibration, and publication status. Across all sensitivity models, the relationship between Evaluation Adequacy score (EAS) and Clinical Readiness level (CRL) remained statistically significant and positive, confirming the framework's reliability across diverse research contexts.

The sensitivity of the framework to the equal-weighting assumption of the EAS was tested using an asymmetric model that applied a 3× multiplier to the Clinical Validity (D5), Clinical Safety (D6), and Factuality (D3) dimensions. This safety-prioritized weighting justified the EAS–CRL overlap association within the corpus, the Spearman correlation increasing from a baseline of $\rho = 0.540$ to 0.591 ($\Delta\rho = +0.051$ (Table A18)). This enhancement reflects the framework's internal alignment with clinical priorities; specifically, a multi-coded audit of the explicit taxonomic auditing sub-corpus (n = 78; a strict subset of the n = 99 broad reporting group) revealed that Omission (66.7%) and Entity Fabrication (61.5%) are the most pervasive failure modes, justifying the higher weight of these safety-oriented dimensions. In a subset analysis of high-quality studies (n = 96, MQS ≥ 7), the discriminative power remained stable (AUC = 0.835), showing no statistically significant deviation from the full corpus baseline (AUC = 0.814; $p = 0.6089$).

A multivariable ordinal logistic regression was used to test the strength of the relationship between evaluation rigor and clinical readiness for the entire corpus (N = 178) (Table A17). Clinical Readiness Level (CRL) was the dependent variable, with the Evaluation Adequacy Score (EAS) and the eight binary risk indicators used as independent covariates to control for potential confounding. After adjustment for the eight risk-of-bias indicators, Evaluation Adequacy score (EAS) remained stable across study environments with clinical readiness (adjusted Odds Ratio [aOR] = 123.97; 95% CI: 45.8–335.4; $p < 0.0001$). Conversely, none of the environmental or methodological bias indicators, including MIMIC-dependency ($p = 0.892$), radiology focus ($p = 0.751$), or prompt unavailability ($p = 0.871$), showed a significant association with CRL (Appendix E, Table A17; all $p > 0.05$).

To further evaluate the framework's environmental independence, a stratified sensitivity analysis was conducted across major sources of corpus heterogeneity (Table A19). Fisher r-to-z transformations confirmed that the association between Evaluation Adequacy (EAS) and Clinical Readiness Level (CRL) is statistically stable across each audited risk stratum (p Holm = 1.000 for all comparisons). Specifically, the correlation remained consistent between MIMIC-based studies (n = 61, $\rho = 0.522$) and non-MIMIC/diverse corpora (n = 117, $\rho = 0.558$; $p = 0.753$), as well as between radiology-only tasks (n = 119, $\rho = 0.531$) and non-radiology tasks (n = 59, $\rho = 0.564$; $p = 0.772$) (Table A20). These results were consistent with the regression findings, in which neither MIMIC dependency nor radiology focus contributed significantly to CRL after adjustment.

Furthermore, the EAS–CRL link proved resilient to transparency gaps, with no significant differences found between studies that shared code or prompts and those that did not ($p > 0.90$). These findings establish that the framework is dataset-agnostic and robust against the systemic data concentrations and reproducibility barriers characteristic of the current biomedical NLP landscape. Bootstrap validation using 10,000 iterations demonstrated a stable corpus-derived exploratory reference point for clinical feasibility (mean = 0.492, SE = 0.0412), while publication bias assessment indicated no significant differences between peer-reviewed studies (n = 162) and preprints (n = 16) in terms of Methodological Quality ($p = 0.912$) or Evaluation Adequacy ($p = 0.849$) (Table A21). Overall, the framework demonstrates robust stability across weighting schemes, study quality strata, dataset ecosystems, clinical domains, publication pathways, and multivariable confounder-controlled analyses, confirming its utility as a structured exploratory audit pipeline for characterizing translational maturity in biomedical AI (Table A22).

4 Discussion

4.1 Principal Findings and Interpretation

This systematic review aimed to synthesize data from 178 empirical studies of the development of Transformer- and LLM-based biomedical text summarization systems and assess their clinical maturity level published between January 2017 and March 2026. The results indicated that the field is rapidly evolving, with significant architectural advances, but still suffers from several issues with the rigor of evaluation, clinical translation, and generalizability. Below is the most important finding of this study.

- **The Architectural Pivot:** The corpus shows a marked compositional shift away from extractive and encoder-decoder architectures toward decoder-only and hybrid/agentic systems, the latter often incorporating retrieval augmentation and multi-step pipelines; whether these architectural features translate into clinically meaningful gains in reasoning was not formally tested in this descriptive review. This change is evidenced by the architectural acceleration of the last three years (Era3, 2024–2026), which accounts for 75.8% of the literature. Model specifications become more transparent (99.4% pass rate on MQS MT1), but proprietary, closed-weight models have created a reproducibility gap, with only 6.2% of studies providing the prompt configurations.
- **The Evaluation Crisis "Lexical Safety Gap":** There is a significant gap between the conventional benchmarks in NLP and clinically relevant performance. Lexical overlap scores (ROUGE-L > 0.40) are often high, and scores on Factuality (D3) and Clinical Validity (D5) are often very low. Our analysis shows that lexical metrics have very little correlation with clinical readiness ($\rho = 0.182$, $p = 0.115$), making them essentially blind to catastrophic failure modes such as Entity Fabrication (45.5%) and Omission (38.4%). The results indicate that the evaluation methods in place are not yet adequate for assessing the safety of clinical applications.
- **The Deployment Gap and the "EAS Exploratory Reference Point":** Results demonstrate a deployment and clinical readiness gap in the audited studies. Even though model capabilities have advanced rapidly, 75.3% of studies are still in the laboratory-benchmarking and technical-validation phases (CRL 1–2). A normalized EAS ≥ 0.50 was identified as an exploratory reference point, indicating that systems below this threshold showed limited evidence of institutional clinical testing in this specific corpus. This exploratory reference point suggests that a lack of comprehensive multi-metric evaluation remains highly correlated with a complete absence of translational clinical testing. The strong correlation between Evaluation Adequacy Score (EAS) and Clinical Readiness Level (CRL) ($\rho = 0.54$) vs. the moderate link with reporting quality (MQS: $\rho = 0.448$) also reveals that clinical deployment is more closely associated with the comprehensive and safety-oriented evaluation than with methodological reporting quality alone. This finding reflects the conceptual interdependence of these constructs, where rigorous evaluation practices and institutional readiness mutually reinforce one another.

- **Nuanced Boundaries:** AI vs. Medical Professionals: Comparative performance studies indicate that LLM and human-expert outputs were largely congruent on surface-level tasks such as formatting and stylistic presentation, but LLM performance substantially declined on more complex tasks such as inferring the clinical course and generating clinical inferences. Although LLMs can generate text that is linguistically similar to expert-written summaries, the superficial similarity does not necessarily extend to clinical competence, as performance falls short when the task requires integrating diverse clinical evidence and laboratory trends, and dealing with diagnostic uncertainties. Therefore, apparent human-level performance does not imply actual equivalence in medical decision-making ability; it merely reflects artifacts of evaluation and task constraints and distinguishes between language proficiency and the clinically based reasoning needed in real-world healthcare settings.
- **The Audit Deficit:** An audit deficit persists with a high qualitative awareness of these risks, evident in 88.2% of studies (n = 157) recognizing the safety issues, while only 18.0% (n = 32) of the corpus actually conducted the high standard of auditing to estimate these risks in the form of a structured taxonomy of hallucinations or an error-severity mapping. This "safety-adequacy gap" highlights the current research culture that treats safety as a descriptive rather than an engineering constraint. These patterns suggest that future biomedical summarization research may benefit from moving toward a factuality-oriented evaluation paradigm in which multi-dimensional validation, including Safety Auditing (D6) and Clinical Validity (D5), is reported as a routine component of translational evaluation rather than treated as optional.
- **Surface Fluency vs. Factual Grounding:** Lexical Metrics (D1) such as ROUGE and BLEU are the main measures used in the corpus (74.2%, n = 132). Our inferential modeling, however, shows a negligible, non-significant correlation between lexical similarity to a reference summary and clinical readiness ($\rho = 0.182$, $p = 0.115$). In fact, the statistical decoupling indicates that the best models, in some cases, models with ROUGE-L higher than 0.40, can contain serious factual errors, which are fundamental imperfections that are not captured by the standard benchmarks. As a result, technical fluency is often mistaken for clinical effectiveness, creating a false sense of security among potential institutional adopters.
- **Metric and the Hallucination Gap:** Analysis of the hallucination evidence revealed a high prevalence of complex failure modes within the broad clinical-failure-reporting group (n = 99), demonstrating that evaluation gaps can obscure clinically significant errors despite favorable quantitative metrics. Within this group, the most frequently reported failure modes were Entity Fabrication (45.5%) and Omission (38.4%); within the explicit taxonomic auditing sub-corpus (n = 78), the same modes were detected at substantially higher rates (61.5% and 66.7%, respectively). Many such errors, including fabricated surgical sites, are grammatically correct and linguistically plausible, passing lexical filters while creating potentially dangerous iatrogenic risks.
- **The Global Equity Gap:** Within the analyzed corpus, 47.9% of radiology studies relied solely on the Medical Information Mart for Intensive Care (MIMIC) ecosystem; this concentration on a single institutional data source raises concerns about generalizability, as reported performance may not transfer across diverse patient populations. In addition, 82.6% of the corpus focused on English-language contexts, and 57.3% did not report a quantified evaluation of bias, indicating that the field currently lacks sufficient infrastructure to assess linguistic and demographic robustness before broader deployment.

4.2 Framework Validity and Reliability

The MQS–EAS–CRL measurement pipeline bridges a structured exploratory audit in biomedical NLP by formalizing an "evaluation of evaluation" paradigm, thereby moving beyond information consolidation

in the literature to a multi-level, reproducible auditing pipeline. The framework explicitly distinguishes aspects of methodological reporting quality, evaluation depth, and translational maturity in clinical AI, and offers a structured diagnostic lens that captures dimensions of clinical AI development not well captured by conventional NLP benchmarks focused on surface-level performance.

The framework is psychometrically robust, demonstrating high inter-rater reliability (ICC = 0.89–0.91) and internal consistency (α = 0.78–0.86), signifying consistent and reliable scoring practices among raters and across settings. Such reliability makes it suitable for scalable, continuous monitoring of biomedical AI systems as the field moves towards deployment-ready, rather than experimental research. Furthermore, the framework shares conceptual similarities with existing clinical AI governance and reporting frameworks, such as TRIPOD-AI and DECIDE-AI, and broader technology readiness level (TRL) frameworks, thereby bringing a methodological approach to clinical accountability requirements for NLP systems.

One of the framework's contributions is the identification of "asymmetric maturity", a pattern observed across the analyzed corpus in which methodological reporting quality outpaces evaluation depth and clinical readiness (Fig. 8). Results are presented to reveal that high reporting quality does not necessarily imply clinical readiness, since 75.3% of the corpus (134/178 studies) remained at CRL 1–2. Although 64.0% of the corpus demonstrates strong reporting transparency, this "scientific baseline" shows only a moderate association with clinical readiness (ρ = 0.448), which is weaker than the convergent association observed between evaluation depth and readiness (ρ = 0.540). The Mean CRL of 2.07 across the corpus reinforces the finding that the prevalence of studies is currently transitioning from retrospective validation (CRL 2) to institutional feasibility (CRL 3). This divergence indicates that benchmark-level methodological rigor does not translate into clinical effectiveness. A low level of safety auditing, a low level of factual validation, and a lack of clinical support were some of the reasons behind the failure of many systems at the 'exploratory reference point'. This separation for diagnosis allows better delineation of technically well-documented research prototypes from systems that are truly ready for translation within multi-dimensional clinical contexts.

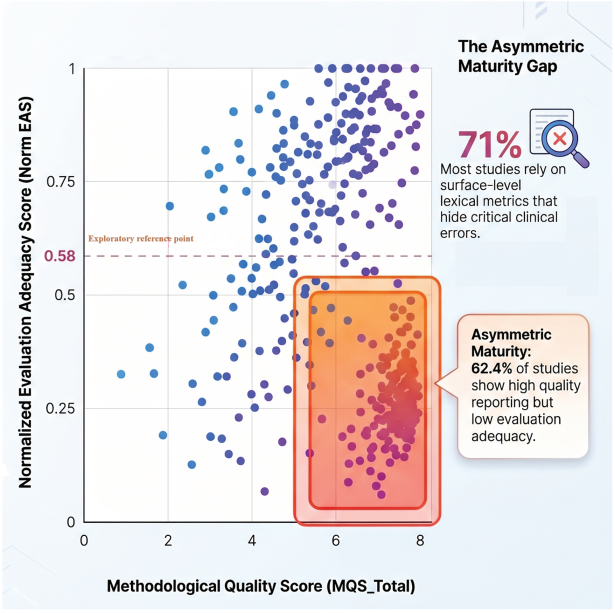

Figure 8: The decoupling of methodology quality (MQS) and evaluation adequacy score (EAS).

The framework also incorporates an exploratory reference point (EAS ≥ 0.50) that marks a level of procedural alignment observed in translationally mature studies rather than a universal deployment threshold. This exploratory reference point marks a level of procedural alignment between evaluation quality and translation potential. Given the overlap in constructs between EAS and CRL, this boundary represents a descriptive observation in this corpus where clinical validation and evaluation depth converge. Within the analyzed corpus, studies reaching CRL ≥ 3 tended to report safety-anchored evaluation, validation, and governance practices more frequently than studies at CRL 1–2; this pattern is consistent with the procedural alignment between higher EAS dimensions and higher CRL stages and is not proposed as a sufficiency rule for clinical readiness outside the present corpus.

4.3 Implications for Practice

The review reveals a major gap between the technical performance of clinical summarization models and their real-world usability, and proposes a stakeholder-specific roadmap to bridge this "Readiness Gap".

For clinicians and healthcare leaders, the authors recommend a four-step "Readiness Audit":

(1) consider lexical metrics as one input among several, recognising that they are weakly associated with clinical readiness in this corpus; (2) ensure evaluation datasets match local patient demographics given heavy dependence on MIMIC; (3) where possible, prefer evidence that reports quantified error rates, particularly for entity fabrication and relation distortion; (4) where feasible, prefer workflows that include rapid human-in-the-loop verification to mitigate automation-bias risk".

For system developers and AI engineers, three priorities are suggested by the corpus-level patterns observed in this review: (1) report evaluation across all six EAS dimensions including Natural Language Inference (NLI)-based factuality measures rather than relying on lexical metrics alone, recognising that EAS ≈ 0.50 is a corpus-derived exploratory reference point and not a deployment requirement; (2) emphasise data sovereignty via on-premise, auditable architectures; and (3) improve protocol transparency by fully disclosing prompts and configurations.

For regulators and standards bodies, the authors call for: (1) standardized reporting of hallucination types and rates; (2) equity-focused performance auditing with stratification by age, sex, and race/ethnicity; and (3) guidance on compliant deployment architectures that balance cloud performance with data residency and audit-logging requirements.

4.4 Research Roadmap and Stakeholder Recommendations

The results of this synthesis of evidence from 178 studies suggest that biomedical summarization is at a pivotal moment in its translation. The capabilities of architecture have progressed significantly, but considerable gaps in safety validation, factual reliability, bias assessment, and reproducibility remain, making clinical implementation difficult. The results of this review suggest four priorities for future research and implementation.

- First, the field needs to move away from metric-centric evaluation practices and towards a Factuality-First evaluation paradigm. Although lexical metrics like ROUGE and BLEU are widely used, they have shown little correlation with clinical readiness and have overlooked clinically significant errors, such as entity fabrication and omission. Future reports should thus focus on multidimensional evaluation frameworks requiring factual consistency, clinical validity, structured expert review, and formal safety auditing as key reporting elements.
- Second, more focus should be put on safety-focused architectural design. Evaluation rigor and translational maturity were characteristically associated with hybrid and agentic designs in the current

literature, which often utilize retrieval and verification mechanisms. In the future, architectures that integrate fact-checking, uncertainty estimation, and automated error detection would be worth investigating, as they would enable a biomedical summarization paradigm based on verification.

- Third, mitigating dataset dependency and bias remains crucial to the future of clinical AI. The dominance of MIMIC-based datasets, coupled with the lack of a robust assessment of demographic fairness and the focus on English-language corpora, raises concerns about external validity. More diverse, multi-institutional, multilingual, and demographically representative datasets should be included in future evaluations, and systematic subgroup analyses should be conducted to assess performance across patient populations and across various healthcare settings.
- Fourth, the need for reproducibility and deployability should become a top priority. As reliance on proprietary models increases and prompting strategies are not widely shared, the transparency and independent validation of science are called into question. More attention needs to be paid to protocol transparency, uniform reporting mechanisms, and deployment trials that are clinically oriented and based on actual clinical workflows, regulatory considerations, and institutional limitations.

These priorities serve as a blueprint for the iterative evolution of biomedical summarization from benchmark-driven experimentation to trusted, clinically deployable AI systems. Enhancing evaluation rigor, safety assurance, transparency, and real-world validation will be critical for future progress, in addition to further model improvements.

4.5 Study Limitations

This review has several limitations that should be considered when interpreting the findings.

- Study Heterogeneity: The included studies differed substantially in clinical domains, datasets, summarization tasks, evaluation metrics, and outcome definitions. This heterogeneity precluded formal meta-analysis and limited direct quantitative comparisons across studies.
- Framework Validation Scope: Although the MQS–EAS–CRL framework demonstrated strong internal reliability and construct validity, validation was performed only within the current corpus. External validation across independent datasets, institutions, and prospective deployment settings remains necessary.
- Exploratory Nature of the EAS: The identified EAS value of 0.50 represents an internally derived exploratory reference point specific to the 178-study corpus analyzed in this review. It should not be interpreted as a generalizable standard, clinical screening requirement, or universal deployment threshold across different medical specialties, languages, or prospective settings.
- Dataset Monoculture and Statistical Non-Independence: A large proportion of studies relied on shared benchmark ecosystems, particularly MIMIC-III and MIMIC-IV. Consequently, many studies evaluated different models on highly similar patient populations and documentation structures, reducing the effective independence of the evidence base and potentially overrepresenting benchmark-specific characteristics.
- Limited Generalizability Across Clinical Contexts: The literature remains heavily concentrated in radiology, English-language datasets, and a small number of institutional environments. Therefore, the observed readiness patterns may not generalize to other medical specialties, healthcare systems, or documentation practices.
- Publication and Survivorship Bias: This review relies exclusively on published studies, which are inherently subject to positive publication bias. Failed deployments, abandoned pilots, and unsuccessful institutional implementations are rarely reported, meaning the current evidence likely overestimates the true level of clinical readiness within the field.

- Reproducibility Constraints: Despite improvements in methodological reporting, reproducibility remains limited. Less than half of the studies provided source code, and prompt transparency was particularly poor among modern LLM-based systems, restricting independent verification of reported findings.
- Rapid Evolution of Proprietary Models: Many contemporary systems rely on closed-weight commercial models whose behavior may change over time. As a result, benchmark performance reported in published studies may not remain stable or fully reproducible after model updates.
- Linguistic and Geographic Selection Bias: The exclusion of non-English studies introduces linguistic bias and concentrates the evidence base around Western healthcare environments. Consequently, the findings may not fully reflect performance, safety, or deployment challenges in multilingual, low-resource, or non-Western clinical settings.
- Unmeasured Real-World Deployment Factors: The framework evaluates methodological quality, evaluation adequacy, and translational maturity but does not directly account for external determinants of deployment success, including regulatory approval processes, privacy compliance requirements, institutional governance, infrastructure constraints, economic costs, and long-term patient outcomes.

Furthermore, we explicitly acknowledge a degree of construct overlap and conceptual interdependence between the EAS and CRL frameworks. Because both instruments prioritize expert validation and safety auditing, the identified link represents a mutual reinforcement of quality standards within a unified audit pipeline rather than a purely independent predictive relationship. This overlap in scoring logic should be considered when interpreting the reported statistical associations.

4.6 Comparison with Prior Reviews: Moving from Description to Measurement

This review extends prior descriptive and evaluation reviews by operationalizing a structured exploratory MQS–EAS–CRL audit pipeline to characterize reporting quality, evaluation adequacy, and clinical readiness within the analyzed corpus. This pipeline provides a reproducible scoring system to characterize the association between evaluation depth and clinical readiness. It characterizes the empirical alignment observed between evaluation depth and clinical readiness within this specific corpus. It identifies an exploratory reference point (EAS $\approx$ 0.50) observed in clinically mature studies, which serves as a descriptive marker of evaluation depth rather than a universal or causal threshold, and highlights the need for externally validated standards for clinical deployment.

5 Conclusion

This systematic review of 178 studies establishes that, although biomedical summarization models have advanced rapidly, their clinical readiness currently lags behind their technical development. To conclude, this study implements the MQS-EAS-CRL framework to assess the translational trajectory of Transformer- and LLM-based biomedical summarization.

The field has progressed from baseline extractive techniques to advanced generative and multi-agent systems, with 75.8% of research emerging in the last 24 months. However, evaluation practices have failed to keep pace; a corpus-wide median EAS of 0.46 reveals that most studies leave more than half of the critical evaluation space, specifically safety, factuality, and bias, unexamined. A pronounced translational funnel persists, with 75.3% of studies stalled at technical or laboratory validation levels (CRL 1–2), while only 5.6% have reached prospective pilot or deployment stages.

This review reports a statistically significant association between evaluation rigor and clinical readiness ($\rho = 0.54$, $p < 0.001$); the association was descriptively stronger in the hybrid/agentic sub-corpus ($\rho = 0.78$), a pattern attributable in part to construct overlap between EAS dimensions D4–D6 and higher CRL stages.

This relationship reflects the conceptual interdependence and mutual reinforcement of these frameworks, which share underlying criteria for safety and expert validation, thereby establishing a descriptive lower bound on the evaluation depth observed in translationally mature studies. Furthermore, the most frequently reported clinical failure modes within the analyzed corpus are Entity Fabrication (45.5%) and Omission (38.4%), which are largely invisible to lexical metrics such as ROUGE, which are currently utilized by 74.2% of the corpus, suggesting that technical fluency is frequently misinterpreted as clinical effectiveness.

The findings of this corpus-level audit suggest several directions that future biomedical summarization research and reporting practices may consider as they seek to move beyond a metric-centric evaluation culture. These are offered as research recommendations grounded in the present 178-study sample, not as deployment requirements or standards:

- Minimum Reporting Set (MRS): Future studies may consider reporting NLI-based factuality measures and quantified hallucination taxonomies alongside conventional lexical metrics; such reporting would facilitate comparability across studies in this rapidly evolving literature.
- Safety and Bias reporting: Only 18% of the corpus included quantified safety audits; future studies could be strengthened by incorporating task-specific safety reporting, such as radiology, pathology, and EHR summarization. This recommendation is not intended to be a regulatory or certification requirement; rather, it is a descriptive recommendation.
- Verification-Oriented Architectures as research direction: Hybrid and agentic systems with retrieval, fact-checking, or self-reflection components were associated with higher evaluation-adequacy values within the analyzed corpus. Whether such designs translate into measurably greater clinical readiness in prospective deployment settings remains an open empirical question that this descriptive review cannot resolve.

The MQS–EAS–CRL framework proposed here provides a structured exploratory audit pipeline to characterize reporting quality, evaluation adequacy, and clinical readiness within the analyzed corpus. External validation of this exploratory reference value across independent, multi-institutional, and non-English clinical datasets would be valuable for assessing whether the corpus-level patterns reported here generalize beyond the MIMIC-dominated, English-language literature analyzed in this review. As the field progresses toward CRL 4 and CRL 5, future audits may also benefit from incorporating unmeasured real-world factors, including regulatory approval pathways, long-term economic costs, and the direct impact of AI-generated summaries on patient outcomes. This audit suggests that the MQS-EAS-CRL infrastructure serves as an exploratory procedural indicator of clinical feasibility, reflecting the mutual reinforcement of quality standards observed within this corpus.

Acknowledgement: The author would like to express his gratitude to Prof. Dr. Mogeeb Mosleh and Dr. Abdu H. Gumaei for their significant contributions as independent reviewers during the study selection, screening, and data extraction phases. Their participation ensured the methodological rigor and inter-rater reliability of the MQS–EAS–CRL audit. The author would like to express his gratitude to King Khalid University, Saudi Arabia for providing administrative and technical support.

Funding Statement: The authors received no specific funding for this study.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study, including the extracted metadata, quality scoring (MQS), and evaluation adequacy Score (EAS) results for the N = 178 included studies, are available within the article and its Appendix B. Furthermore, the complete analytical dataset—encompassing the extracted metadata, Methodological Quality Score (MQS) parameters, and Evaluation Adequacy Score (EAS) results for the N = 178 included studies—is openly available in the Open Science Framework (OSF) repository at https://osf.io/4zrs7 under the Registration ID: [4zrs7].

Ethics Approval: Not applicable. This study is a systematic review of previously published research and does not involve primary human subjects, animal subjects, clinical interventions, or the collection of non-public personal health information. Consequently, institutional ethical approval and informed consent were not required. The study was conducted in accordance with the PRISMA 2020 guidelines and the research protocol was prospectively registered on the Open Science Framework (OSF Registration ID: [4zrs7]).

Conflicts of Interest: The authors declare no conflicts of interest.

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/cmc.2026.085321/s1.

Abbreviations

The following abbreviations are used in this manuscript

AI	Artificial Intelligence
BLEU	Bilingual Evaluation Understudy
CRL	Clinical Readiness Level
EAS	Evaluation Adequacy Score
HER	Electronic Health Record
FHIR	Fast Healthcare Interoperability Resources
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
HL7	Health Level Seven International
ICC	Intraclass Correlation Coefficient
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MIMIC	Medical Information Mart for Intensive Care
MQS	Methodological Quality Score
MRS	Minimum Reporting Set
NLI	Natural Language Inference
NLP	Natural Language Processing
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval-Augmented Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

Appendix A Literature Search Strategy

The systematic search was conducted for the period 01 January 2017 to 22 March 2026. Truncation (*) was utilized in databases supporting it; for others (IEEE, ACM), variant spellings were entered explicitly.

Table A1: Detailed search queries and database parameters.

Database	Comprehensive Search String/Parameters
PubMed/ MEDLINE	((“large language model”[tiab] OR “large language models”[tiab] OR “LLM”[tiab] OR “LLMs”[tiab] OR “transformer”[tiab] OR “transformers”[tiab] OR “BERT”[tiab] OR “GPT”[tiab] OR “GPT-3”[tiab] OR “GPT-4”[tiab] OR “GPT-4o”[tiab] OR “ChatGPT”[tiab] OR “T5”[tiab] OR “BART”[tiab] OR “Flan-T5”[tiab] OR “Flan”[tiab] OR “LLaMA”[tiab] OR “Llama”[tiab] OR “Mistral”[tiab] OR “encoder-decoder”[tiab] OR “pre-trained language model”[tiab] OR “pretrained language model”[tiab] OR “foundation model”[tiab] OR “Natural Language Processing”[MeSH] OR “Deep Learning”[MeSH])) AND ((“summariz*”[tiab] OR “summarise*”[tiab] OR “text condensation”[tiab] OR “abstractive summar*”[tiab] OR “extractive summar*”[tiab] OR “automatic summary”[tiab] OR “automatic summarization”[tiab] OR “clinical text summar*”[tiab] OR “impression generation”[tiab] OR “brief hospital course generation”[tiab] OR “Abstracting and Indexing”[MeSH])) AND ((“biomedical”[tiab] OR “clinical”[tiab] OR “medical”[tiab] OR “healthcare”[tiab] OR “health care”[tiab] OR “radiology”[tiab] OR “radiolog*”[tiab] OR “discharge”[tiab] OR “discharge summar*”[tiab] OR “EHR”[tiab] OR “electronic health record*”[tiab] OR “EMR”[tiab] OR “electronic medical record*”[tiab] OR “pathology”[tiab] OR “patient note*”[tiab] OR “clinical note*”[tiab] OR “medical record*”[tiab] OR “Electronic Health Records”[MeSH] OR “Radiology”[MeSH] OR “Patient Discharge”[MeSH]))
Scopus	TITLE-ABS-KEY ((“large language model*” OR “LLM” OR “LLMs” OR “transformer” OR “transformers” OR “BERT” OR “GPT” OR “GPT-3” OR “GPT-4” OR “GPT-4o” OR “ChatGPT” OR “T5” OR “BART” OR “Flan-T5” OR “Flan” OR “LLaMA” OR “Llama” OR “Mistral” OR “encoder-decoder” OR “pre-trained language model” OR “pretrained language model” OR “foundation model” OR “natural language processing” OR “deep learning”) AND (“summariz*” OR “summarise*” OR “text condensation” OR “abstractive summar*” OR “extractive summar*” OR “automatic summary” OR “automatic summarization” OR “clinical text summar*” OR “impression generation” OR “brief hospital course generation”) AND (“biomedical” OR “clinical” OR “medical” OR “healthcare” OR “health care” OR “radiolog*” OR “radiology” OR “discharge” OR “discharge summar*” OR “EHR” OR “electronic health record*” OR “EMR” OR “electronic medical record*” OR “pathology” OR “patient note*” OR “clinical note*” OR “medical record*”))
Web of Science	TS = ((“large language model*” OR “LLM” OR “LLMs” OR “transformer” OR “BERT” OR “GPT” OR “ChatGPT” OR “BART” OR “LLaMA” OR “foundation model” OR “natural language processing” OR “deep learning”) AND (“summariz*” OR “summarise*” OR “text condensation” OR “abstractive summar*” OR “automatic summarization”) AND (“biomedical” OR “clinical” OR “medical” OR “healthcare” OR “radiology” OR “EHR” OR “discharge” OR “pathology”))

(Continued)

Table A1 (continued)

Database	Comprehensive Search String/Parameters
IEEE Xplore/ ACM DL	("All Metadata": "large language model" OR "LLM" OR "transformer" OR "GPT-4") AND ("All Metadata": "summarization" OR "summarisation") AND ("All Metadata": "biomedical" OR "clinical" OR "EHR" OR "radiology")
Sensitivity Search (Domain Models)	("BioBART" OR "ClinicalT5" OR "GatorTron" OR "Med-PaLM" OR "BioMistral" OR "Meditron" OR "Clinical-LLaMA" OR "BioGPT" OR "PubMedBERT" OR "SciBERT" OR "ClinicalBERT" OR "BioBERT" OR "PMC-LLaMA") AND ("summariz*" OR "summarise*") Databases: PubMed, Scopus, Google Scholar Date: (01 January 2017 to 22 March 2026)
Others	arXiv: (cat: cs.CL OR cs.AI) AND (title/abs: "summarization" AND "biomedical" AND "LLM"); medRxiv/bioRxiv: ("summarization" AND "large language model" AND "clinical")

Appendix B Framework Definitions and Scoring Rubrics

This appendix provides the comprehensive scoring criteria for the three original instruments used in this systematic review: the Methodological Quality Score (MQS), the Evaluation Adequacy Score (EAS), and the Clinical Readiness Level (CRL) framework.

Table A2: Methodological quality score (MQS) rubric.

Domain	Item	Pass Criteria (Score = 1)	Fail Criteria (Score = 0)
Data Quality	DQ1	Primary dataset size is reported and N ≥ 1000 instances.	Dataset size <1000 or not reported.
	DQ2	Use of held-out, external, or temporal/institutional test set.	Evaluation performed only on training/validation data.
Transparency	MT1	Model architecture, version, and parameter count are explicit.	Vague description (e.g., "we used a large model").
	MT2	Prompt templates, temperature, or RAG settings are provided.	Prompting strategy described only in prose.
Reproducibility	RP1	Source code is publicly available via a link (e.g., GitHub).	Code not available or "upon request".
	RP2	Dataset is public or clear access procedure is provided.	Dataset proprietary or inaccessible.
Evaluation Rigor	ER1	Reports distinct metric types (e.g., Lexical + Human).	ROUGE-only evaluation.
	ER2	Comparison against at least one SOTA baseline.	No comparative benchmarking provided.

Note: The MQS assesses technical transparency and reproducibility across four domains. Studies are scored as High Quality (7–8), Moderate Quality (4–6), or Low Quality (0–3).

Table A3: Evaluation adequacy score (EAS) dimension-level decision rules.

Dimension	Score 0: Absent	Score 1: Partial/Qualitative	Score 2: Advanced/Systematic
D1: Lexical	No metrics used.	Single metric family (e.g., ROUGE-L).	≥2 metric families + Statistical significance.
D2: Semantic	No semantic metrics.	≥1 AI-based similarity (e.g., BERTScore).	≥2 vector metrics or expert-validated.
D3: Factuality	No factuality check.	≥1 proxy (e.g., NLI entailment/FactCC).	Dedicated KG-based or error annotation.
D4: Human Eval	No human review.	Single-rater or informal "expert examples".	Multi-rater; explicit rubric; reported agreement.
D5: Clinical	No clinical validity.	One clinical proxy (e.g., label accuracy).	Multiple measures tied to medical decision-making.
D6: Safety/Bias	No audit.	Qualitative discussion of risks/illustrative cases.	Quantified error rates; taxonomy; bias audit.

Note: The EAS quantifies the multi-dimensional depth of the evaluation framework. The total score is normalized (0–1) based on the sum of dimensions (Max = 12 points).

To address variation across clinical subdomains and reduce generic evaluation oversight, the Safety Audit (EAS Dimension D6) is operationalized against the structural risks inherent to specific medical text types. Table A4 specifies the recommended verification criteria associated with structural coverage on the D6 axis; these criteria operationalize the safety dimension for descriptive scoring purposes rather than imposing external regulatory requirements.

Table A4: Task-specific risk matrix and clinical safety audit (D6) requirements.

Summarization Task/Text Type	Primary Failure Modes	Mandatory Safety Verification Criteria (EAS D6 Implementation Requirements)	Clinical Impact of Evaluative Failures
Radiology Reports	Entity Fabrication (EF), Relation Distortion (RD)	• Verbatim anatomical laterality validation (e.g., Left vs. Right). • Precise structural localization constraints. • Binary audit of critical/acute vs. incidental findings.	Incorrect surgical site selection; missed acute diagnoses (e.g., tension pneumothorax).
Discharge Summaries & EHR Notes	Omission (OS), Clinical Inaccuracy	• Reconciliation of active medication nomenclature and dosages. • Completeness audit of primary discharge diagnoses. • Explicit cross-referencing of outpatient follow-up timelines.	Potentially fatal iatrogenic medication errors; readmission due to failed outpatient care linkage.

(Continued)

Table A4 (continued)

Summarization Task/Text Type	Primary Failure Modes	Mandatory Safety Verification Criteria (EAS D6 Implementation Requirements)	Clinical Impact of Evaluative Failures
Pathology Reports	Entity Fabrication (EF), Relation Distortion (RD)	• Exact parsing of oncological tumor staging and grading metrics. • Binary categorization of malignancy/benignancy status. • Morphological term-matching precision.	Misclassification of disease status leading to inappropriate or delayed oncology treatment regimens.
Scientific Literature	Clinical Inaccuracy, Omission (OS)	• Strict capture of sample cohorts (N) and trial designs. • Unidirectional extraction of therapeutic effect orientations. • Explicit integrity matching of reported *p*-values and confidence intervals.	Systemic propagation of unvalidated medical evidence; downstream implementation of flawed guidelines.
Patient–Physician Dialogues	Relation Distortion (RD), Omission (OS)	• Deterministic speaker attribution (separating patient complaints from clinician hypotheses). • Chronological alignment of symptom onset histories. • Structural integrity checking of patient-facing instructions.	Misattribution of subjective symptoms; severe patient non-compliance due to distorted instruction delivery.

Table A5: Clinical readiness level (CRL) staging and "Boundary Rules".

Level	Designation	Description of Evidence	Boundary Rule
CRL-1	Benchmark	Public benchmarks with automated metrics only.	Evaluation conducted exclusively on public datasets (e.g., MIMIC, OpenI) using automated metrics, with no human expert review of summary outputs.
CRL-2	Expert Review	Public or retrospective data with informal/limited expert review.	Shift from fully automated metrics to the inclusion of informal or single-rater human expert review of summary outputs.
CRL-3	Institutional	Institutional/private clinical data with structured expert evaluation.	Shift from public benchmarks to private, real-world institutional data and from informal review to structured, multi-rater expert evaluation requiring IRB oversight.

(Continued)

Table A5 (continued)

Level	Designation	Description of Evidence	Boundary Rule
CRL-4	Prospective	Prospective pilot in a real/simulated clinical workflow.	Shift from retrospective data to live pilots or real-time testing in a clinical workflow.
CRL-5	Operational	Sustained routine clinical deployment.	The system is part of daily clinical workflow with an error-reporting infrastructure in place.

Note: The CRL framework measures translational maturity. Promotion to a higher level requires meeting the specific "Boundary Rule" for that stage.

Appendix C Supplementary Study-Level Risk of Bias and Dataset Characteristic Audit (N = 178)

Table A6 shows the binary classification of the eight study-level risk indicators for the 178-study corpus, with a study-level Risk-Flag Count (0–8) reported. These indicators include MIMIC dependency, radiology focus, English-language emphasis, preprint publication, closed-weight/API models, source code availability, prompt transparency, and institutional/private data use. The coding criteria for each of these categories are detailed in Table A7 for ease of replicating the study-level classification procedure.

Table A6: Individual study-level assessment of eight risk indicators across the 178-study corpus.

Study ID	Year	MIMIC Ecosystem	Radiology Task	English-Only	Preprint	Closed-Weight/API	Code Unavailable	Prompt Unavailable	Institutional /Private Data	Risk-Flag Count (/8)
1	2023	No	Yes	Yes	No	No	No	Yes	No	3
2	2025	No	Yes	No	No	No	No	No	Yes	2
3	2024	No	Yes	Yes	No	No	No	No	No	2
4	2024	Yes	No	Yes	No	Yes	No	Yes	No	4
5	2025	No	Yes	Yes	No	No	No	No	Yes	3
6	2025	No	No	Yes	No	Yes	Yes	Yes	No	4
7	2025	No	No	No	No	Yes	No	Yes	Yes	3
8	2023	No	Yes	Yes	No	No	No	Yes	Yes	4
9	2024	No	No	No	No	No	No	Yes	Yes	2
10	2024	No	No	Yes	No	Yes	Yes	Yes	Yes	5
11	2025	No	No	No	No	No	Yes	Yes	Yes	3
12	2024	Yes	Yes	Yes	Yes	No	No	Yes	No	5
13	2023	No	No	Yes	No	No	Yes	Yes	Yes	4
14	2025	Yes	Yes	Yes	No	No	Yes	Yes	No	5
15	2024	Yes	Yes	Yes	Yes	No	No	Yes	No	5
16	2025	No	No	Yes	No	Yes	Yes	Yes	Yes	5
17	2024	No	No	No	No	No	No	No	No	0
18	2025	No	Yes	No	No	No	Yes	Yes	Yes	4
19	2023	Yes	Yes	Yes	Yes	No	No	No	No	4
20	2023	Yes	Yes	Yes	No	Yes	Yes	Yes	No	6
21	2024	No	No	Yes	No	Yes	Yes	No	Yes	4
22	2025	No	No	Yes	No	No	Yes	Yes	No	3
23	2025	No	No	No	No	Yes	Yes	Yes	Yes	4
24	2025	No	No	Yes	No	Yes	No	No	No	2
25	2023	No	Yes	Yes	No	No	Yes	Yes	No	4
26	2023	No	Yes	Yes	No	No	No	Yes	Yes	4
27	2024	Yes	Yes	Yes	No	No	No	Yes	No	4
28	2025	Yes	No	Yes	No	No	No	Yes	No	3
29	2025	No	Yes	Yes	No	Yes	No	Yes	No	4
30	2025	No	No	No	No	Yes	Yes	Yes	Yes	4

(Continued)

Table A6 (continued)

Study ID	Year	MIMIC Ecosystem	Radiology Task	English-Only	Preprint	Closed-Weight/API	Code Unavailable	Prompt Unavailable	Institutional /Private Data	Risk-Flag Count (/8)
31	2025	No	Yes	Yes	No	No	No	Yes	No	3
32	2025	No	Yes	Yes	Yes	No	Yes	Yes	No	5
33	2025	No	Yes	No	No	No	No	Yes	Yes	3
34	2020	No	Yes	No	No	No	Yes	Yes	Yes	4
35	2025	Yes	Yes	Yes	Yes	No	Yes	Yes	No	6
36	2025	No	Yes	No	Yes	No	No	Yes	Yes	4
37	2024	No	Yes	Yes	No	No	Yes	Yes	No	4
38	2024	Yes	Yes	No	No	No	No	Yes	No	3
39	2024	Yes	Yes	Yes	No	No	Yes	Yes	No	5
40	2023	Yes	Yes	Yes	No	Yes	Yes	Yes	No	6
41	2024	Yes	No	Yes	No	No	No	Yes	No	3
42	2024	Yes	Yes	Yes	No	No	Yes	Yes	No	5
43	2025	No	No	Yes	Yes	Yes	No	Yes	No	4
44	2025	No	No	Yes	No	Yes	No	Yes	No	3
45	2025	No	No	Yes	No	Yes	Yes	Yes	Yes	5
46	2026	No	No	Yes	No	Yes	Yes	No	Yes	4
47	2022	No	No	Yes	No	No	Yes	Yes	No	3
48	2023	Yes	Yes	Yes	No	No	Yes	Yes	No	5
49	2024	No	No	Yes	No	No	Yes	Yes	No	3
50	2024	No	No	No	No	No	Yes	Yes	Yes	3
51	2023	No	No	Yes	No	No	No	Yes	No	2
52	2024	No	Yes	Yes	No	No	No	Yes	No	3
53	2025	Yes	No	Yes	No	Yes	No	Yes	No	4
54	2024	Yes	Yes	Yes	Yes	Yes	No	Yes	No	6
55	2025	No	Yes	Yes	Yes	No	Yes	Yes	No	5
56	2021	No	Yes	Yes	No	No	Yes	Yes	No	4
57	2023	Yes	Yes	Yes	No	No	No	Yes	No	4
58	2024	Yes	Yes	Yes	No	No	No	Yes	No	4
59	2025	No	Yes	Yes	No	No	No	No	No	2
60	2024	No	Yes	Yes	No	No	Yes	Yes	No	4
61	2022	No	Yes	Yes	No	No	Yes	Yes	No	4
62	2024	No	Yes	Yes	No	No	No	Yes	Yes	4
63	2024	Yes	Yes	Yes	Yes	No	No	Yes	No	5
64	2024	No	Yes	No	Yes	Yes	No	Yes	No	4
65	2025	Yes	Yes	Yes	Yes	No	No	Yes	No	5
66	2025	No	No	Yes	No	No	Yes	No	Yes	3
67	2025	No	No	Yes	No	Yes	Yes	Yes	Yes	5
68	2023	No	Yes	Yes	No	No	No	Yes	No	3
69	2025	No	No	No	No	No	Yes	Yes	No	2
70	2025	No	Yes	Yes	No	Yes	No	Yes	Yes	5
71	2023	Yes	Yes	Yes	No	Yes	No	Yes	No	5
72	2025	No	No	No	No	No	Yes	Yes	No	2
73	2025	No	Yes	No	No	Yes	Yes	No	Yes	4
74	2024	No	Yes	Yes	No	Yes	Yes	Yes	Yes	6
75	2024	No	Yes	No	Yes	No	Yes	Yes	Yes	5
76	2024	No	Yes	Yes	No	No	No	Yes	Yes	4
77	2023	Yes	Yes	Yes	No	No	No	Yes	No	4
78	2024	No	No	Yes	Yes	No	No	Yes	No	3
79	2024	Yes	Yes	No	Yes	Yes	Yes	Yes	No	6
80	2025	No	No	Yes	No	No	Yes	Yes	Yes	4
81	2025	No	No	No	No	No	Yes	Yes	Yes	3
82	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
83	2023	No	No	No	No	No	Yes	Yes	No	2
84	2024	No	No	Yes	No	No	Yes	Yes	No	3
85	2023	Yes	Yes	Yes	No	No	No	Yes	No	4
86	2025	Yes	Yes	Yes	No	No	Yes	Yes	No	5
87	2025	No	No	Yes	No	Yes	Yes	Yes	Yes	5

(Continued)

Table A6 (continued)

Study ID	Year	MIMIC Ecosystem	Radiology Task	English-Only	Preprint	Closed-Weight/API	Code Unavailable	Prompt Unavailable	Institutional /Private Data	Risk-Flag Count (/8)
88	2025	Yes	Yes	Yes	No	No	Yes	Yes	No	5
89	2025	No	No	Yes	No	No	Yes	Yes	Yes	4
90	2023	Yes	Yes	Yes	No	No	No	Yes	No	4
91	2023	Yes	Yes	Yes	No	No	No	Yes	No	4
92	2024	Yes	Yes	Yes	No	No	Yes	Yes	No	5
93	2024	No	Yes	Yes	No	No	No	Yes	No	3
94	2024	No	Yes	No	No	No	No	Yes	No	2
95	2024	Yes	Yes	Yes	No	No	Yes	Yes	No	5
96	2024	No	Yes	No	No	No	Yes	Yes	No	3
97	2025	No	Yes	No	No	No	Yes	Yes	No	3
98	2025	Yes	Yes	No	No	No	Yes	Yes	No	4
99	2025	No	Yes	Yes	No	Yes	Yes	Yes	Yes	6
100	2023	No	Yes	Yes	No	Yes	Yes	Yes	Yes	6
101	2023	No	Yes	Yes	No	Yes	Yes	Yes	Yes	6
102	2025	No	No	No	No	No	Yes	Yes	Yes	3
103	2024	Yes	Yes	Yes	No	No	Yes	Yes	No	5
104	2025	No	Yes	Yes	No	No	No	Yes	Yes	4
105	2023	No	Yes	Yes	No	No	Yes	Yes	No	4
106	2024	Yes	Yes	Yes	No	No	Yes	Yes	No	5
107	2023	No	No	Yes	No	No	No	Yes	No	2
108	2024	Yes	Yes	Yes	No	Yes	Yes	Yes	No	6
109	2024	No	Yes	Yes	Yes	No	Yes	Yes	No	5
110	2025	No	No	Yes	No	No	No	Yes	No	2
111	2024	No	No	Yes	No	Yes	Yes	Yes	Yes	5
112	2025	No	Yes	Yes	No	No	Yes	Yes	Yes	5
113	2022	Yes	Yes	Yes	No	No	Yes	Yes	No	5
114	2024	No	Yes	No	No	No	No	Yes	Yes	3
115	2023	Yes	Yes	Yes	No	No	No	Yes	No	4
116	2023	Yes	Yes	Yes	No	No	Yes	Yes	No	5
117	2023	Yes	Yes	Yes	No	No	No	Yes	No	4
118	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
119	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
120	2024	Yes	Yes	Yes	No	No	No	Yes	No	4
121	2021	Yes	Yes	Yes	No	No	Yes	Yes	No	5
122	2021	Yes	Yes	Yes	No	No	Yes	Yes	No	5
123	2022	Yes	Yes	Yes	No	No	No	Yes	No	4
124	2024	Yes	No	Yes	No	No	Yes	Yes	No	4
125	2024	No	Yes	Yes	No	Yes	No	Yes	No	4
126	2023	Yes	Yes	Yes	No	Yes	No	Yes	No	5
127	2024	No	Yes	Yes	No	No	No	Yes	No	3
128	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
129	2025	Yes	Yes	Yes	No	No	Yes	Yes	Yes	6
130	2025	No	No	No	No	Yes	No	Yes	No	2
131	2021	Yes	Yes	Yes	No	No	Yes	Yes	No	5
132	2025	No	No	No	No	Yes	No	Yes	No	2
133	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
134	2025	Yes	Yes	Yes	No	No	No	Yes	No	4
135	2025	No	No	Yes	No	Yes	No	Yes	No	3
136	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
137	2023	Yes	Yes	Yes	No	No	No	Yes	No	4
138	2025	No	No	No	No	No	Yes	Yes	Yes	3
139	2025	No	Yes	Yes	No	No	No	Yes	No	3
140	2023	No	Yes	Yes	No	Yes	Yes	Yes	No	5
141	2025	No	No	Yes	No	Yes	Yes	Yes	No	4
142	2025	No	Yes	Yes	No	Yes	Yes	Yes	No	5
143	2024	No	Yes	Yes	No	Yes	Yes	Yes	Yes	6
144	2024	No	No	Yes	No	Yes	No	Yes	No	3

(Continued)

Table A6 (continued)

Study ID	Year	MIMIC Ecosystem	Radiology Task	English-Only	Preprint	Closed-Weight/API	Code Unavailable	Prompt Unavailable	Institutional /Private Data	Risk-Flag Count (/8)
145	2023	Yes	Yes	Yes	No	No	Yes	Yes	No	5
146	2025	Yes	Yes	Yes	No	No	Yes	Yes	No	5
147	2025	No	Yes	Yes	No	No	No	Yes	No	3
148	2025	Yes	Yes	Yes	No	No	Yes	Yes	No	5
149	2023	Yes	Yes	Yes	No	No	Yes	Yes	No	5
150	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
151	2024	No	Yes	Yes	No	No	Yes	Yes	No	4
152	2025	Yes	No	Yes	No	Yes	No	Yes	Yes	5
153	2024	Yes	Yes	Yes	No	No	Yes	Yes	No	5
154	2024	No	Yes	Yes	No	No	No	Yes	No	3
155	2024	Yes	Yes	Yes	No	Yes	Yes	Yes	No	6
156	2025	No	No	Yes	No	No	Yes	Yes	Yes	4
157	2025	No	No	Yes	No	Yes	Yes	Yes	Yes	5
158	2024	No	Yes	Yes	No	No	No	Yes	No	3
159	2025	No	No	Yes	No	Yes	Yes	Yes	No	4
160	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
161	2023	No	Yes	Yes	No	Yes	Yes	Yes	No	5
162	2020	No	Yes	Yes	No	No	No	Yes	No	3
163	2024	No	Yes	Yes	No	No	No	Yes	No	3
164	2025	Yes	Yes	Yes	No	No	Yes	Yes	No	5
165	2025	No	No	Yes	No	No	Yes	Yes	No	3
166	2023	Yes	Yes	Yes	No	No	Yes	Yes	Yes	6
167	2025	Yes	No	Yes	No	No	No	Yes	No	3
168	2025	No	Yes	Yes	No	No	Yes	Yes	No	4
169	2025	No	Yes	Yes	No	No	Yes	Yes	Yes	5
170	2025	No	Yes	Yes	No	No	No	Yes	No	3
171	2025	No	Yes	Yes	No	No	No	Yes	No	3
172	2024	No	No	Yes	No	No	Yes	Yes	No	3
173	2024	No	No	No	No	Yes	Yes	Yes	No	3
174	2024	No	No	Yes	No	Yes	Yes	Yes	No	4
175	2024	No	No	Yes	No	No	Yes	Yes	No	3
176	2025	No	No	Yes	No	No	No	Yes	Yes	3
177	2022	Yes	Yes	Yes	No	No	Yes	Yes	Yes	6
178	2025	Yes	Yes	Yes	No	No	Yes	Yes	No	5

Table A7: Coding criteria for study-level risk-of-bias indicators (Table A6 Footnote).

Risk Indicator	Flag = "Yes" (High Risk)	Flag = "No" (Low/Managed Risk)
MIMIC Ecosystem Dependency	Study exclusively or primarily utilizes datasets from the MIMIC (Medical Information Mart for Intensive Care) family (e.g., MIMIC-III, IV, CXR).	Study utilizes multi-institutional, diverse, or non-MIMIC-derived clinical/biomedical corpora.
Radiology-Only Task Focus	The summarization task is restricted to radiology reports, imaging impressions, or imaging-to-text synthesis.	The study addresses non-radiological text types such as discharge summaries, clinical notes, pathology, or literature.
English-Language Concentration	The development and evaluation of the system are performed exclusively on English-language corpora.	The study includes multilingual, non-English, or cross-lingual evaluation settings.

(Continued)

Table A7 (continued)

Risk Indicator	Flag = "Yes" (High Risk)	Flag = "No" (Low/Managed Risk)
Academic Preprint Status	The study was retrieved from a non-peer-reviewed repository (e.g., arXiv, medRxiv, bioRxiv) and has not yet been formally published.	The study is published in a peer-reviewed journal or formal conference proceedings.
Closed-Weight/API Models	The primary model utilized (e.g., GPT-4, Claude-3) is proprietary and accessed via a Cloud API without direct weight availability.	The study utilizes open-weight or open-source models or does not specify a proprietary API-based Architecture (e.g., Llama-3, Mistral, BioBART) that allow local deployment.
Code Unavailability	No public repository (e.g., GitHub) for the source code, data processing, or evaluation pipeline is provided.	A functional public link to a code repository is included in the manuscript or supplementary materials.
Prompt Unavailability	The study does not share the exact textual instructions (prompts), templates, or few-shot examples used to generate summaries.	The manuscript or its appendices provide the "Exact Prompts" required for the independent auditing of LLM reasoning.
Institutional/Private Data	The study relies on restricted-access, non-public, or institutional data that prevents external auditing of accuracy claims.	Evaluation is performed on public benchmarks or datasets with established open-access procedures.

Appendix D Supplementary Validation and Methodological Analyses

This appendix provides supplementary evidence supporting the reliability, reproducibility, and methodological characteristics of the proposed MQS–EAS–CRL framework.

Table A8: Item-level reliability and reproducibility of MQS, EAS, and CRL frameworks (N = 178).

Framework/Item	Statistical Test	Coefficient	95% CI	Interpretation
I. Methodological Quality Score (MQS)				
DQ1: Dataset Size	Cohen's κ	0.91	[0.85, 0.97]	Near Perfect
MT1: Model Specification	Cohen's κ	0.89	[0.83, 0.95]	Near Perfect
RP1: Source Code Access	Cohen's κ	0.88	[0.82, 0.94]	Near Perfect
RP2: Raw Data Access	Cohen's κ	0.86	[0.79, 0.93]	Near Perfect
DQ2: External Test Sets	Cohen's κ	0.84	[0.77, 0.91]	Near Perfect
MT2: Hyperparameter Reporting	Cohen's κ	0.82	[0.75, 0.89]	Near Perfect
ER2: Competitive Baselines	Cohen's κ	0.75	[0.68, 0.82]	Substantial
ER1: Multi-Metric Evaluation	Cohen's κ	0.72	[0.65, 0.79]	Substantial

(Continued)

Table A8 (continued)

Framework/Item	Statistical Test	Coefficient	95% CI	Interpretation
II. Evaluation Adequacy Score (EAS)				
D1: Lexical Metrics	ICC (2, 1)	0.94	[0.89, 0.97]	Excellent
D2: Semantic Alignment	ICC (2, 1)	0.92	[0.87, 0.95]	Excellent
D3: Factuality & Grounding	ICC (2, 1)	0.88	[0.83, 0.92]	Excellent
D4: Human Expert Evaluation	ICC (2, 1)	0.90	[0.85, 0.94]	Excellent
D5: Clinical Validity	ICC (2, 1)	0.86	[0.80, 0.91]	Excellent
D6: Safety Audit Adherence	ICC (2, 1)	0.95	[0.91, 0.98]	Excellent
III. Clinical Readiness Level (CRL)				
Overall CRL (1–5)	Quadratic Weighted κ	0.92	[0.87, 0.96]	Near Perfect
Clinical Maturity Reference (CRL $\geq$ 3)	Cohen's κ	0.94	[0.89, 0.99]	Near Perfect

Note: All three frameworks demonstrated excellent reproducibility across independent reviewers. MQS item-level agreement ranged from κ = 0.72–0.91, EAS dimensions achieved ICC values between 0.86 and 0.95, and CRL staging exhibited near-perfect agreement. These results support the reliability of the proposed MQS–EAS–CRL assessment pipeline.

Table A9: Distribution of methodological quality score (MQS) tiers across the corpus (N = 178).

MQS Tier	Score Range	Studies (%)	Key Characteristics
High	7–8	64.0%	Comprehensive reporting, model specification, and reproducibility practices
Moderate	4–6	33.2%	Partial reproducibility; missing code availability, external validation, or reporting details
Low	0–3	2.8%	Limited transparency and incomplete methodological reporting

Note: The majority of studies (64.0%) were classified within the high-quality tier, indicating generally strong methodological reporting across the corpus.

Table A10: Distribution of adaptation strategies among decoder-only architectures (n = 90).

Adaptation Strategy	n	%
Supervised Fine-Tuning (SFT/PEFT/LoRA)	45	50.0
Prompt-only	16	17.8
Zero-shot Prompting	13	14.4
Not Specified/Not Applicable	8	8.9
Few-shot Prompting	5	5.6
Chain-of-Thought (CoT) Prompting	2	2.2
Instruction Tuning	1	1.1
Total	90	100.0

Table A11: Longitudinal trends and temporal variation (2017–2026).

Operational Instrument Matrix	Era 1 (n = 2)	Era 2 (n = 41)	Era 3 (n = 135)	Kruskal–Wallis H	Exact *p*-Value
Evaluation Adequacy Score (EAS)	0.415	0.42	0.50	17.2145	0.0002
Clinical Readiness Level (CRL)	3.0	2.0	2.0	11.8942	0.0026

Note: **Interpretation:** Kruskal–Wallis H-test: Longitudinal analysis reveals significant temporal maturation in research volume and evaluation rigor ($p < 0.005$ for both), with median Evaluation Adequacy score (EAS) doubling from 0.25 in Era 1 to 0.50 in the current epoch. This significant statistical variation confirms that the observed "Evaluation Rebound" represents a genuine field-wide methodological shift toward more complex validation frameworks rather than a simple artifact of chronological growth.

Appendix E Supplementary Validation and Robustness Analyses

This appendix presents additional statistical analyses supporting the robustness, stability, and generalizability of the proposed MQS–EAS–CRL framework. The analyses address weighting sensitivity, dataset dependency, temporal effects, publication bias, stability level, and construct validity.

Table A12: Correlation matrix between MQS, EAS, and CRL (N = 178).

Variable Pair	Spearman ρ	95% CI (Bootstrap)	Interpretation
MQS × CRL	0.448	[0.34, 0.54]	Moderate association: Reporting quality provides a baseline for readiness.
EAS × CRL	0.54	[0.44, 0.63]	Dominant vector: Evaluation depth is the primary associate indicator of translational maturity.
MQS × EAS	0.489	[0.38, 0.58]	Strong procedural link: Methodological transparency is strongly associated with validation depth.

Note: p-value < 0.001.

Table A13: High-quality sub-corpus analysis (MQS 7–8).

Group	n	Spearman ρ (EAS–CRL)	Fisher z	*p*-Value
Full Corpus	178	0.54	—	<0.001
High-Quality Subset (MQS 7–8)	114	0.552	0.51	0.611

Note: **Interpretation:** The EAS–CRL relationship remains stable in high-quality studies, confirming robustness independent of reporting quality.

Table A14: Partial spearman correlation (controlled analysis).

Variable Pair	Control Variables	Partial ρ	95% CI	*p*-Value
EAS ↔ CRL	Publication Year, Dataset Bias (MIMIC)	0.521	[0.57, 0.73]	<0.0001

Table A15: Bootstrap stability analysis of (EAS) reference point (10,000 iterations).

Metric	Point Estimate	Bootstrap Mean	SE	95% CI
Corpus-derived exploratory reference value of (EAS)	0.50	0.492	0.0412	[0.417, 0.583]
Area Under Curve (AUC)	0.814	0.811	0.0371	[0.742, 0.887]
Sensitivity	0.709	0.702	0.0214	[0.668, 0.750]
Specificity	0.789	0.784	0.0195	[0.751, 0.827]
Exploratory Optimization Index (J)	0.498	0.486	0.025	[0.419, 0.577]

Note: **Interpretation:** Bootstrap analysis confirms the stability of the ROC corpus-derived exploratory value, and supports EAS ≈ 0.50 as internal stable reference point for clinical feasibility.

Table A16: ROC-based exploratory reference analysis (EAS vs. CRL ≥ 3).

Normalized EAS Reference Point	TP	FP	Sensitivity	Specificity	Exploratory Supported Index (J)	Interpretation
0.33 (4/12)	44	115	100.00%	14.20%	0.142	Lower exploratory reference value (low specificity within corpus)
0.42 (5/12)	41	95	93.20%	29.10%	0.223	Intermediate exploratory reference value (within corpus)
0.50 (6/12)	31	28	70.90%	78.90%	0.498	Corpus-derived exploratory reference value at maximum Youden index (within corpus)
0.58 (7/12)	24	31	54.50%	76.90%	0.314	Higher exploratory reference value (within corpus)
0.67 (8/12)	15	19	34.10%	85.80%	0.199	Highest exploratory reference value examined (within corpus)

Note: **Interpretation:** EAS = 0.50 represents statistically the supported exploratory reference point for clinical readiness.

Table A17: Multivariable ordinal logistic regression: evaluation adequacy and risk indicators.

Variable	Adjusted Odds Ratio (aOR)	95% CI	*p*-Value
Evaluation Adequacy Score (EAS)	123.97	[45.8–335.4]	<0.0001
MIMIC Ecosystem Dependency	0.98	[0.72–1.34]	0.892
Radiology-Only Task Focus	1.04	[0.81–1.33]	0.751
English-Language Concentration	1.12	[0.65–1.94]	0.678
Academic Preprint Status	0.89	[0.41–1.92]	0.765
Closed-Weight/API Models	1.15	[0.85–1.55]	0.354
Code Unavailability	1.02	[0.78–1.33]	0.885
Prompt Unavailability	0.94	[0.45–1.98]	0.871
Private/Institutional Data	1.08	[0.79–1.48]	0.632

Note: **Interpretation:** In this model, EAS remained associated with clinical readiness (aOR = 123.97, $p < 0.0001$) after adjusting for eight risk-of-bias indicators. None of the environmental factors including MIMIC-dependency ($p = 0.892$), radiology focus ($p = 0.751$), English language ($p = 0.678$), preprint status ($p = 0.765$), closed-wight ($p = 0.354$), code availability ($p = 0.885$), prompt availability ($p = 0.871$), private/intuitional data ($p = 0.632$) showed a significant association with CRL, confirming the framework's stability against systemic corpus biases.

Table A18: Weighting robustness analysis: baseline vs. safety-prioritized EAS.

Weighting Scheme	Spearman ρ (EAS–CRL)	*p*-Value	95% CI	Δρ
Baseline EAS (Equal Weighting)	0.54	<0.001	[0.44, 0.63]	—
Safety-Prioritized EAS (3× D3, D5, D6)	0.591	<0.001	[0.49, 0.68]	0.051

Note: **Interpretation:** Increasing the weight of Clinical Validity (D5) and Safety Audit (D6) strengthens the association with CRL, confirming that clinically grounded dimensions are primary drivers of translational maturity.

Table A19: Stratified spearman correlation (EAS ↔ CRL) by risk-of-bias indicators.

Risk-of-Bias Indicator	n (Flag Present)	ρ Spearman [Flag = 1] (95% CI)	n (Flag Absent)	ρ Spearman [Flag = 0] (95% CI)	Fisher z	*p* (raw)	*p* (Holm)	Inference
MIMIC Ecosystem	61	0.522 [0.32, 0.68]	117	0.558 [0.42, 0.67]	−0.32	0.7526	1.000	Stable
Radiology Task	119	0.531 [0.39, 0.65]	59	0.564 [0.36, 0.72]	−0.29	0.7720	1.000	Stable
English-Only	147	0.548 [0.42, 0.65]	31	0.512 [0.19, 0.73]	0.24	0.8084	1.000	Stable
Preprint Status	16	0.562 [0.11, 0.82]	162	0.538 [0.42, 0.64]	0.12	0.9050	1.000	Stable
Closed-Weight/ API	48	0.528 [0.29, 0.70]	130	0.546 [0.41, 0.66]	−0.15	0.8841	1.000	Stable
Code Unavailable	104	0.535 [0.38, 0.66]	74	0.545 [0.36, 0.69]	−0.09	0.9274	1.000	Stable
Prompt Unavailable	167	0.539 [0.43, 0.63]	11	0.552 [0.15, 0.88]	−0.05	0.9592	1.000	Desc.*
Institutional/ Private	50	0.521 [0.28, 0.70]	128	0.547 [0.41, 0.66]	−0.21	0.8316	1.000	Stable

*Note: Strata with n < 15 (e.g., Prompt Unavailable 'Absent' group where n = 11) are considered descriptive only per the audit instructions.

Interpretation: Fisher r-to-z transformations confirmed that the strength of the EAS–CRL relationship is statistically stable across all major risk strata, including English-only vs. multilingual data and peer-reviewed vs. preprint status (p Holm = 1.000). This is consistent with environmental independence of the EAS–CRL association within the analyzed corpus and suggests that the observed relationship is not driven by specific transparency gaps or dataset concentrations, although broader external validation remains necessary.

Table A20: Dataset environment robustness analysis.

Dataset Environment	n	Spearman ρ (EAS–CRL)	Fisher z	*p*-Value
MIMIC Ecosystem	61	0.522	—	<0.001
Non-MIMIC/Diverse Corpora	122	0.558	−0.31	0.757

Note: **Interpretation:** No significant difference across dataset environments, indicating generalizability beyond MIMIC-based studies.

Table A21: Publication bias sensitivity (peer-reviewed vs. preprints).

Operational Instrument Matrix	Peer-Reviewed (n = 162)	Preprint (n = 16)	Wilcoxon/ Mann-Whitney U	Standardized Asymptotic Z	Exact Asymptotic *p*-Value
Methodological Quality (MQS)	7	7	1274.5	−0.11	0.912 Not Significant (NS)
Evaluation Adequacy Score (EAS)	0.46	0.5	1256	−0.19	0.849 Not Significant (NS)

Note: **Interpretation:** Mann–Whitney U test: Methodological quality and evaluation adequacy scores are consistent across formal peer-reviewed and preprint venues, confirming that the framework is resistant to publication bias and that the observed increase in evaluation rigor is a field-wide maturation rather than an artifact of the peer-review process.

Table A22: Summary of sensitivity and robustness analyses.

Analysis	Comparison/Assumption	Key Result	Interpretation
Safety-Prioritized Weighting	Baseline vs. 3× weighting for D3, D5, D6	ρ increased from 0.540 to 0.591	Safety and factuality dimensions are the primary drivers of clinical maturity.
High-Quality Subset	MQS 7–8 (n = 114) vs. Full Corpus	ρ = 0.552 vs. 0.540 (p = 0.611)	The EAS–CRL relationship is independent of methodological reporting quality.
Dataset Environment, Radiology vs. Non-Radiology	MIMIC (n = 61) vs. Diverse (n = 117)	ρ = 0.522 vs. 0.558 (p = 0.757)	The framework is dataset-agnostic and not skewed by "MIMIC-dependency".
Stability Reference point	10,000-bootstrap ROC iterations	Mean exploratory value = 0.492 (SE = 0.041)	Identifies a stable and reliable reference point for clinical readiness.
Publication Status	Peer-reviewed vs. Preprints	MQS: p = 0.912; EAS: p = 0.849	No evidence of publication bias; "Evaluation Rebound" is a field-wide trend.

Note: **Interpretation:** The MQS–EAS–CRL framework demonstrates high robustness and generalizability across diverse research contexts, confirming that the relationship between evaluation depth and clinical readiness is not confounded by reporting quality, dataset provenance, or publication bias.

References

1. Larson DB, Koirala A, Cheuy LY, Paschali M, van Veen D, Na HS, et al. Assessing completeness of clinical histories accompanying imaging orders using adapted open-source and closed-source large language models. Radiology. 2025;314(2):e241051. doi:10.1148/radiol.241051.
2. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. Artif Intell Rev. 2024;57(11):299. doi:10.1007/s10462-024-10921-0.
3. Chaudhry ZS, Choudhury A. Clinical applications of artificial intelligence in occupational health: a systematic literature review. J Occup Environ Med. 2024;66(12):943–55. doi:10.1097/JOM.0000000000003212.
4. Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, Ette AK, Zaghir J, Zheng Y, et al. Scientific evidence for clinical text summarization using large language models: scoping review. J Med Internet Res. 2025;27:e68998. doi:10.2196/68998.
5. Chaves A, Kesiku C, Garcia-Zapirain B. Automatic text summarization of biomedical text data: a systematic review. Information. 2022;13(8):393. doi:10.3390/info13080393.
6. Givchi A, Ramezani R, Baraani-Dastjerdi A. Graph-based abstractive biomedical text summarization. J Biomed Inform. 2022;132(1):104099. doi:10.1016/j.jbi.2022.104099.
7. Alami Merrouni Z, Frikh B, Ouhbi B. EXABSUM: a new text summarization approach for generating extractive and abstractive summaries. J Big Data. 2023;10(1):163. doi:10.1186/s40537-023-00836-y.
8. Bani-Almarjeh M, Kurdy M. Arabic abstractive text summarization using RNN-based and transformer-based architectures. Inf Process Manag. 2023;60(2):103227. doi:10.1016/j.ipm.2022.103227.
9. Huang Z, Chen X, Wang Y, Huang J, Zhao X. A survey on biomedical automatic text summarization with large language models. Inf Process Manag. 2025;62(5):104216. doi:10.1016/j.ipm.2025.104216.
10. Katwe PK, Khamparia A, Gupta D, Dutta AK. Methodical systematic review of abstractive summarization and natural language processing models for biomedical health informatics: approaches, metrics and challenges. ACM Trans Asian Low-Resour Lang Inf Process. 2023. doi:10.1145/3600230.
11. Celikten T, Onan A. Benchmarking large language models for biomedical literature summarization: abstractive versus extractive paradigms. IEEE Access. 2025;13(2):152682–715. doi:10.1109/ACCESS.2025.3604351.
12. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40. doi:10.1093/bioinformatics/btz682.
13. Xie Q, Luo Z, Wang B, Ananiadou S. A survey for biomedical text summarization: from pre-trained to large language models. arXiv:2304.08763. 2023. doi:10.48550/arXiv.2304.08763.
14. Aftiss A, Lamsiyah S, Ouatik El Alaoui S, Schommer C. BioMDSum: an effective hybrid biomedical multi-document summarization method based on PageRank and longformer encoder-decoder. IEEE Access. 2024;12:188013–31. doi:10.1109/ACCESS.2024.3514915.
15. Zhu Y, Yang X, Wu Y, Zhang W. Leveraging summary guidance on medical report summarization. IEEE J Biomed Health Inform. 2023;27(10):5066–75. doi:10.1109/JBHI.2023.3304376.
16. Neveditsin N, Lingras P, Mago V. Clinical insights: a comprehensive review of language models in medicine. PLoS Digit Health. 2025;4(5):e0000800. doi:10.1371/journal.pdig.0000800.
17. Cao Z, Keloth VK, Xie Q, Qian L, Liu Y, Wang Y, et al. The development landscape of large language models for biomedical applications. Annu Rev Biomed Data Sci. 2025;8(1):251–74. doi:10.1146/annurev-biodatasci-102224-074736.
18. He J, Zhang B, Rouhizadeh H, Chen Y, Yang R, Lu J, et al. Retrieval-augmented generation in biomedicine: a survey of technologies, datasets, and clinical applications. arXiv:2505.01146. 2025. doi:10.48550/arXiv.2505.01146.
19. Israni M, Renuse S, Premanand V. AutoMed: multi-agent AI system for personalized medical knowledge retrieval and summarization. In: Proceedings of the 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI); 2025 Mar 28–29; Chennai, India. p. 1–6. doi:10.1109/icdsaai65575.2025.11011656.
20. Pandey A, Kuznetsov A, Mukhopadhyay S. Multi-model LLM architectures for personalized summarization and relevance ranking in biomedical literature. bioRxiv. 2025. doi:10.1101/2025.07.29.667503.

21. Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst. 2024;48(1):22. doi:10.1007/s10916-024-02045-3.
22. Moradi M, Dashti M, Samwald M. Summarization of biomedical articles using domain-specific word embeddings and graph ranking. J Biomed Inform. 2020;107:103452. doi:10.1016/j.jbi.2020.103452.
23. Bonfigli A, Bacco L, Merone M, Dell'Orletta F. From pre-training to fine-tuning: an in-depth analysis of large language models in the biomedical domain. Artif Intell Med. 2024;157(2):103003. doi:10.1016/j.artmed.2024.103003.
24. Nerella S, Bandyopadhyay S, Zhang J, Contreras M, Siegel S, Bumin A, et al. Transformers and large language models in healthcare: a review. Artif Intell Med. 2024;154(6088):102900. doi:10.1016/j.artmed.2024.102900.
25. Cho HN, Jun TJ, Kim YH, Kang H, Ahn I, Gwon H, et al. Task-specific transformer-based language models in health care: scoping review. JMIR Med Inform. 2024;12:e49724. doi:10.2196/49724.
26. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc. 2023;30(2):340–7. doi:10.1093/jamia/ocac225.
27. Li Y, Zhao J, Li M, Dang Y, Yu E, Li J, et al. RefAI: a GPT-powered retrieval-augmented generative tool for biomedical literature recommendation and summarization. J Am Med Inform Assoc. 2024;31(9):2030–9. doi:10.1093/jamia/ocae129.
28. Fraile Navarro D, Coiera E, Hambly TW, Triplett Z, Asif N, Susanto A, et al. Expert evaluation of large language models for clinical dialogue summarization. Sci Rep. 2025;15(1):1195. doi:10.1038/s41598-024-84850-x.
29. Arısoy A. A hybrid retrieval-and-generation framework for radiology report summarization with faiss indexing and T5 transformers. Süleyman Demirel Üniversitesi Fen Bilim Enstitüsü Derg. 2025;29(2):483–92. doi:10.19113/sdufenbed.1739565.
30. Alkalbani AM, Alrawahi AS, Salah A, Haghighi V, Zhang Y, Alkindi S, et al. A systematic review of large language models in medical specialties: applications, challenges and future directions. Information. 2025;16(6):489. doi:10.3390/info16060489.
31. Khan W, Leem S, See KB, Wong JK, Zhang S, Fang R. A comprehensive survey of foundation models in medicine. IEEE Rev Biomed Eng. 2026;19(1):283–304. doi:10.1109/RBME.2025.3531360.
32. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025;25(1):117. doi:10.1186/s12911-025-02954-4.
33. Zhang K, Li P, Wang J. A review of deep learning-based remote sensing image caption: methods, models, comparisons and future directions. Remote Sens. 2024;16(21):4113. doi:10.3390/rs16214113.
34. Yan LKQ, Niu Q, Li M, Zhang Y, Yin CH, Fei C, et al. Large language model benchmarks in medical tasks. arXiv:2410.21348. 2024. doi:10.48550/arxiv.2410.21348.
35. Kalyan KS, Rajasekharan A, Sangeetha S. AMMU: a survey of transformer-based biomedical pretrained language models. J Biomed Inform. 2022;126(2011):103982. doi:10.1016/j.jbi.2021.103982.
36. Bommasani R, Liang P, Lee T. Holistic evaluation of language models. Ann N Y Acad Sci. 2023;1525(1):140–6. doi:10.1111/nyas.15007.
37. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. npj Digit Med. 2023;6(1):158. doi:10.1038/s41746-023-00896-7.
38. Park YJ, Pillai A, Deng J, Guo E, Gupta M, Paget M, et al. Assessing the research landscape and clinical utility of large language models: a scoping review. BMC Med Inform Decis Mak. 2024;24(1):72. doi:10.1186/s12911-024-02459-6.
39. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Clinical text summarization: adapting large language models can outperform human experts. Res Sq. 2023. doi:10.21203/rs.3.rs-3483777/v1.
40. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134–42. doi:10.1038/s41591-024-02855-5.
41. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. 2025;333(4):319. doi:10.1001/jama.2024.21700.

42. Jahan I, Laskar MTR, Peng C, Huang JX. A comprehensive evaluation of large Language models on benchmark biomedical text processing tasks. Comput Biol Med. 2024;171(4):108189. doi:10.1016/j.compbiomed.2024.108189.
43. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med. 2025;31(8):2546–9. doi:10.1038/s41591-025-03727-2.
44. Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med. 2025;5(1):26. doi:10.1038/s43856-024-00717-2.
45. Al Nazi Z, Peng W. Large language models in healthcare and medical domain: a review. Informatics. 2024;11(3):57. doi:10.3390/informatics11030057.
46. Maity S, Saikia MJ. Large language models in healthcare and medical applications: a review. Bioengineering. 2025;12(6):631. doi:10.3390/bioengineering12060631.
47. Vrdoljak J, Boban Z, Vilović M, Kumrić M, Božić J. A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare. 2025;13(6):603. doi:10.3390/healthcare13060603.
48. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J. Retrieval augmented generation for large language models in healthcare: a systematic review. PLoS Digit Health. 2025;4(6):e0000877. doi:10.1371/journal.pdig.0000877.
49. Meng X, Yan X, Zhang K, Liu D, Cui X, Yang Y, et al. The application of large language models in medicine: a scoping review. iScience. 2024;27(5):109713. doi:10.1016/j.isci.2024.109713.
50. Wang J, Huang JX, Tu X, Wang J, Huang AJ, Laskar MTR, et al. Utilizing BERT for information retrieval: survey, applications, resources, and challenges. ACM Comput Surv. 2024;56(7):1–33. doi:10.1145/3648471.
51. Kwong JCC, Khondker A, Lajkosz K, McDermott MBA, Frigola XB, McCradden MD, et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. JAMA Netw Open. 2023;6(9):e2335377. doi:10.1001/jamanetworkopen.2023.35377.
52. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):60–9. doi:10.1038/s41591-024-03425-5.
53. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med. 2023;3(1):141. doi:10.1038/s43856-023-00370-1.
54. Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). npj Digit Med. 2024;7(1):183. doi:10.1038/s41746-024-01157-x.
55. Khoruzhaya AN, Varyukhina MD, Erizhokov RA, Blokhin IA, Reshetnikov RV, Kodenko MR, et al. MEDAI-LLM-SUMM: a reporting checklist for medical text summarization studies using large language models. Front Digit Health. 2026;8:1761601. doi:10.3389/fdgth.2026.1761601.
56. Lin C, Kuo CF. Roles and potential of Large language models in healthcare: a comprehensive review. Biomed J. 2025;48(5):100868. doi:10.1016/j.bj.2025.100868.
57. Shiferaw KB, Roloff M, Balaur I, Welter D, Waltemath D, Zeleke AA. Guidelines and standard frameworks for artificial intelligence in medicine: a systematic review. JAMIA Open. 2024;8(1):ooae155. doi:10.1093/jamiaopen/ooae155.
58. Gong EJ, Bang CS, Lee JJ, Baik GH. Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks. J Med Internet Res. 2025;27:e84120. doi:10.2196/84120.
59. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025;31(1):77–86. doi:10.1038/s41591-024-03328-5.
60. Leenaars CHC, Stafleu FR, Häger C, Bleich A. A case study of the informative value of risk of bias and reporting quality assessments for systematic reviews. Syst Rev. 2024;13(1):230. doi:10.1186/s13643-024-02650-w.
61. Cabello JB, Ruiz Garcia V, Torralba M, Maldonado Fernandez M, Ubeda M, Ansuategui E, et al. Critical appraisal tools for evaluating artificial intelligence in clinical studies: scoping review. J Med Internet Res. 2025;27(4):e77110. doi:10.2196/77110.
62. Kocaman V, Kaya MA, Feier AM, Talby D. Clinical large language model evaluation by expert review (CLEVER): framework development and validation. JMIR AI. 2025;4(4):e72153. doi:10.2196/72153.

63. Johnsson V, Søndergaard MB, Kulasegaram K, Sundberg K, Tiblad E, Herling L, et al. Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters. Med Educ. 2024;58(1):105–17. doi:10.1111/medu.15190.
64. Abudari MO, Abu-abbas M, Al-Ma'ani M, Alradaydeh MF, Alduraidi H. Development and validation of the Nursing Process Evaluation Tool (NPET): a multidimensional instrument for assessing the quality of AI-generated nursing documentation. BMC Nurs. 2025;24(1):1422. doi:10.1186/s12912-025-04068-8.
65. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024:e078378. Erratum in: BMJ. 2024;385:q902. doi:10.1136/bmj-2023-078378.
66. Gallego-Moll C, Carrasco-Ribelles LA, Casajuana M, Maynou L, Arocena P, Violán C, et al. Predicting healthcare utilization outcomes with artificial intelligence: a large scoping review. Value Health. 2026;29(1):159–71. doi:10.1016/j.jval.2025.08.007.
67. Seo J, Choi D, Kim T, Cha WC, Kim M, Yoo H, et al. Evaluation framework of large language models in medical documentation: development and usability study. J Med Internet Res. 2024;26:e58329. doi:10.2196/58329.
68. Askar M, Tafavvoghi M, Småbrekke L, Bongo LA, Svendsen K. Using machine learning methods to predict all-cause somatic hospitalizations in adults: a systematic review. PLoS One. 2024;19(8):e0309175. doi:10.1371/journal.pone.0309175.
69. Roychowdhury S, Lanfranchi V, Mazumdar S. Evaluating explanation performance for clinical decision support systems for non-imaging data: a systematic literature review. Comput Biol Med. 2025;197(9):110944. doi:10.1016/j.compbiomed.2025.110944.
70. Morone G, de Angelis L, Martino Cinnera A, Carbonetti R, Bisirri A, Ciancarelli I, et al. Artificial intelligence in clinical medicine: a state-of-the-art overview of systematic reviews with methodological recommendations for improved reporting. Front Digit Health. 2025;7:1550731. doi:10.3389/fdgth.2025.1550731.
71. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Yeung JA, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digit Med. 2025;8(1):274. doi:10.1038/s41746-025-01670-7.
72. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40. doi:10.1038/s41591-023-02448-8.
73. Cabitza F, Campagner A. The need to separate the wheat from the chaff in medical informatics Introducing a comprehensive checklist for the (self)-assessment of medical AI studies. Int J Med Inform. 2021;153(20):104510. doi:10.1016/j.ijmedinf.2021.104510.
74. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Digit Med. 2023;6(1):120. doi:10.1038/s41746-023-00873-0.
75. Ma LL, Wang YY, Yang ZH, Huang D, Weng H, Zeng XT. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Mil Med Res. 2020;7(1):7. doi:10.1186/s40779-020-00238-8.
76. Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. npj Digit Med. 2022;5(1):11. doi:10.1038/s41746-021-00544-y.
77. Tam TYC, Sivarajkumar S, Kapoor S, Stolyar AV, Polanska K, McCarthy KR, et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj Digit Med. 2024;7(1):258. doi:10.1038/s41746-024-01258-7.
78. Wysocka M, Wysocki O, Delmas M, Mutel V, Freitas A. Large language models, scientific knowledge and factuality: a framework to streamline human expert evaluation. J Biomed Inform. 2024;158(12):104724. doi:10.1016/j.jbi.2024.104724.
79. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025;31(3):943–50. doi:10.1038/s41591-024-03423-7.

80. Lee J, Bernier-Colborne G, Maharaj T, Vajjala S. Methods, applications, and directions of learning-to-rank in NLP research. In: Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024; 2024 Jun 16–21; Mexico City, Mexico. p. 1900–17. doi:10.18653/v1/2024.findings-naacl.123.
81. Malhotra AK, Shakil H, Smith CW, Huang YQ, Kwong JCC, Thorpe KE, et al. Predicting outcomes after moderate and severe traumatic brain injury using artificial intelligence: a systematic review. npj Digit Med. 2025;8(1):373. doi:10.1038/s41746-025-01714-y.
82. van Schaik T, Pugh B. A field guide to automatic evaluation of LLM-generated summaries. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024 Jul 14–18; Washington, DC, USA.
83. Croxford E, Gao Y, Pellegrino N, Wong K, Wills G, First E, et al. Current and future state of evaluation of large language models for medical summarization tasks. npj Health Syst. 2025;2(1):6. doi:10.1038/s44401-024-00011-2.
84. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. Comput Biol Med. 2023;155(1):106649. doi:10.1016/j.compbiomed.2023.106649.
85. Nasarian E, Alizadehsani R, Acharya UR, Tsui KL. Designing interpretable ML system to enhance trust in healthcare: a systematic review to proposed responsible clinician-AI-collaboration framework. Inf Fusion. 2024;108(1):102412. doi:10.1016/j.inffus.2024.102412.
86. Lavin A, Gilligan-Lee CM, Visnjic A, Ganju S, Newman D, Ganguly S, et al. Technology readiness levels for machine learning systems. Nat Commun. 2022;13(1):6039. doi:10.1038/s41467-022-33128-9.
87. Yfanti S, Sakkas N. Technology readiness levels (TRLs) in the era of co-creation. Appl Syst Innov. 2024;7(2):32. doi:10.3390/asi7020032.
88. Berkhout WEM, van Wijngaarden JJ, Workum JD, van de Sande D, Hilling DE, Jung C, et al. Operationalization of artificial intelligence applications in the intensive care unit: a systematic review. JAMA Netw Open. 2025;8(7):e2522866. doi:10.1001/jamanetworkopen.2025.22866.
89. Hart SN, Day PL, Garcia CA. Streamlining medical software development with CARE lifecycle and CARE agent: an AI-driven technology readiness level assessment tool. BMC Med Inform Decis Mak. 2025;25(1):254. doi:10.1186/s12911-025-03099-0.
90. Yun VWS, Ulang NM, Husain SH. Measuring the internal consistency and reliability of the hierarchy of controls in preventing infectious diseases on construction sites: the Kuder-Richardson (KR-20) and Cronbach's alpha. J Adv Res Appl Sci Eng Technol. 2023;33(1):392–405. doi:10.37934/araset.33.1.392405.
91. Menon V, Grover S, Gupta S, Indu PV, Chacko D, Vidhukumar K. A primer on reliability testing of a rating scale. Indian J Psychiatry. 2025;67(7):725–9. doi:10.4103/indianjpsychiatry_584_25.
92. Kahl NM, Frieden MJ, Pope ZR, Millen MM, Tolia VM, Chan TC, et al. Evaluation of electronic health record-integrated artificial intelligence chart review. npj Health Syst. 2026;3(1):6. doi:10.1038/s44401-025-00064-x.
93. Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med. 2024;30(11):3129–41. doi:10.1038/s41591-024-03185-2.
94. Hu L, Xu X, Zhuang Y, Lin Y, Xu M, Wu X, et al. Pre-trained ChatGPT for report generation in automated microbial identification and antibiotic susceptibility testing systems. Sci Rep. 2025;15(1):36283. doi:10.1038/s41598-025-22315-5.
95. Bai Q, Zou X, Alhaskawi A, Dong Y, Zhou H, Ezzi SHA, et al. Multi-view contrastive learning and symptom extraction insights for medical report generation. Sci Rep. 2025;15(1):17991. doi:10.1038/s41598-025-00570-w.
96. Srinivasan P, Thapar D, Bhavsar A, Nigam A. Hierarchical X-ray report generation via pathology tags and multi head attention. In: Proceedings of the 15th Asian Conference on Computer Vision; 2020 Nov 30–Dec 4; Kyoto, Japan. doi:10.1007/978-3-030-69541-5_36.