Review

Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review

Lydie Bednarczyk[1*], MD; Daniel Reichenpfader[2,3*], MSc; Christophe Gaudet-Blavignac[1*], PhD; Amon Kenna Ette[1,3], BSc; Jamil Zaghir[1,3], MSc; Yuanyuan Zheng[1,3], MSc; Adel Bensahla[1,3], MSc; Mina Bjelogrlic[1,3], PhD; Christian Lovis[1,3], MPH, MD

[1]Division of Medical Information Sciences, University Hospital of Geneva, Geneva, Switzerland

[2]Institute for Patient-centered Digital Health, Bern University of Applied Sciences, Biel, Switzerland

[3]Faculty of Medicine, University of Geneva, Geneva, Switzerland

[*]these authors contributed equally

Corresponding Author:
Lydie Bednarczyk, MD
Division of Medical Information Sciences
University Hospital of Geneva
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 797686970
Email: lydie.bednarczyk@hug.ch

Abstract

Background: Information overload in electronic health records requires effective solutions to alleviate clinicians' administrative tasks. Automatically summarizing clinical text has gained significant attention with the rise of large language models. While individual studies show optimism, a structured overview of the research landscape is lacking.

Objective: This study aims to present the current state of the art on clinical text summarization using large language models, evaluate the level of evidence in existing research and assess the applicability of performance findings in clinical settings.

Methods: This scoping review complied with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. Literature published between January 1, 2019, and June 18, 2024, was identified from 5 databases: PubMed, Embase, Web of Science, IEEE Xplore, and ACM Digital Library. Studies were excluded if they did not describe transformer-based models, did not focus on clinical text summarization, did not engage with free-text data, were not original research, were nonretrievable, were not peer-reviewed, or were not in English, French, Spanish, or German. Data related to study context and characteristics, scope of research, and evaluation methodologies were systematically collected and analyzed by 3 authors independently.

Results: A total of 30 original studies were included in the analysis. All used observational retrospective designs, mainly using real patient data (n=28, 93%). The research landscape demonstrated a narrow research focus, often centered on summarizing radiology reports (n=17, 57%), primarily involving data from the intensive care unit (n=15, 50%) of US-based institutions (n=19, 73%), in English (n=26, 87%). This focus aligned with the frequent reliance on the open-source Medical Information Mart for Intensive Care dataset (n=15, 50%). Summarization methodologies predominantly involved abstractive approaches (n=17, 57%) on single-document inputs (n=4, 13%) with unstructured data (n=13, 43%), yet reporting on methodological details remained inconsistent across studies. Model selection involved both open-source models (n=26, 87%) and proprietary models (n=7, 23%). Evaluation frameworks were highly heterogeneous. All studies conducted internal validation, but external validation (n=2, 7%), failure analysis (n=6, 20%), and patient safety risks analysis (n=1, 3%) were infrequent, and none reported bias assessment. Most studies used both automated metrics and human evaluation (n=16, 53%), while 10 (33%) used only automated metrics, and 4 (13%) only human evaluation.

Conclusions: Key barriers hinder the translation of current research into trustworthy, clinically valid applications. Current research remains exploratory and limited in scope, with many applications yet to be explored. Performance assessments often lack reliability, and clinical impact evaluations are insufficient raising concerns about model utility, safety, fairness, and data privacy. Advancing the field requires more robust evaluation frameworks, a broader research scope, and a stronger focus on real-world applicability.

(J Med Internet Res 2025;27:e68998) doi: 10.2196/68998

KEYWORDS

summarization; large language models; natural language processing; health care; electronic health records; scoping review; translational research; artificial intelligence

Introduction

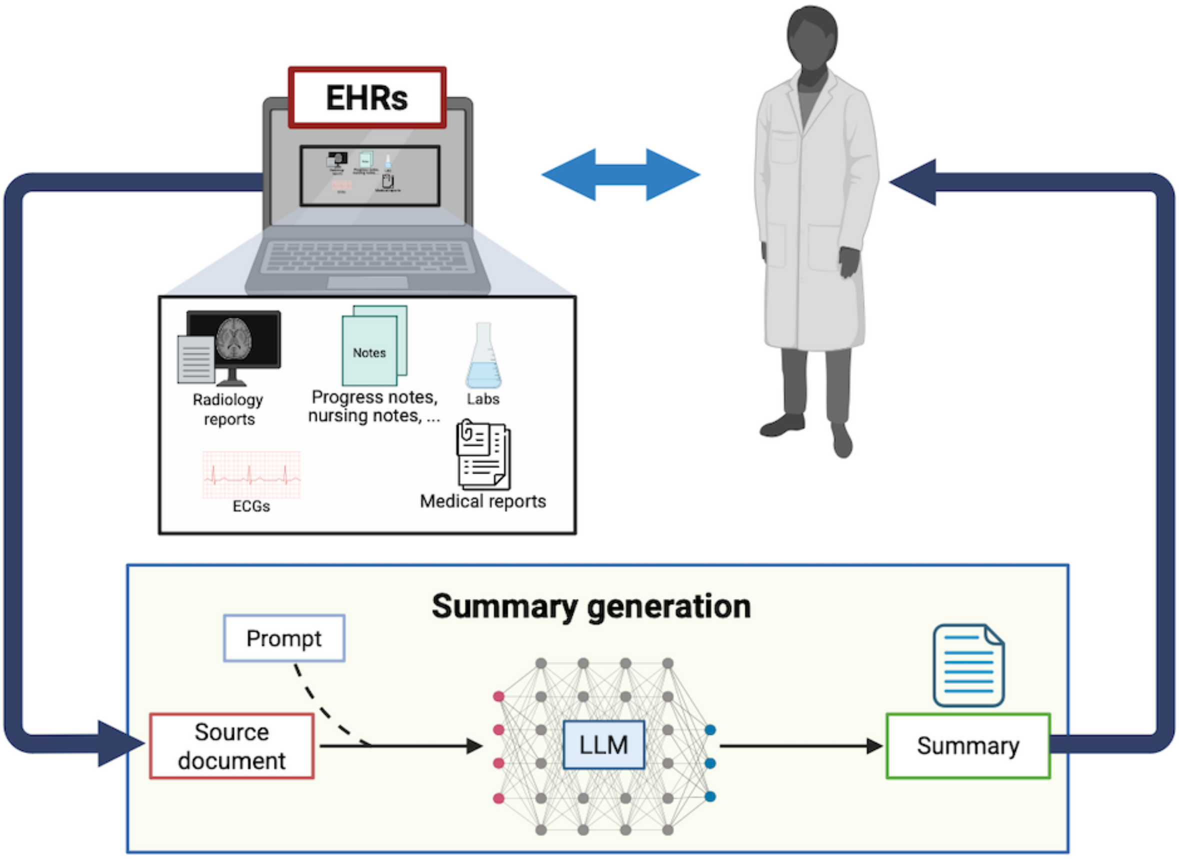
In February 2024, Van Veen et al [1] reported that large language models (LLMs) could outperform medical experts in clinical text summarization. Their work investigated the effectiveness of specifically tailored models to accurately summarize clinical documents. However, a careful analysis of the experimental design and the evaluation methodology questions this statement.

Clinical text summarization is described by Keszthelyi et al [2] as the art of collecting, synthesizing, and communicating patient information. An effective summary must be tailored to meet the needs of its intended audience, which requires a clear definition of the clinical text summary's purpose to ensure relevance and meaning.

In the fast-paced environment of modern health care, coupled with information overload in electronic health records (EHRs), physicians face added cognitive load and time pressure. Misunderstandings, incomplete information sharing, or delays in conveying critical patient details ultimately affect the quality of care and decision-making [3]. Thus, reducing the administrative burden on clinicians has become a critical need.

LLMs can process significant volumes of input data and produce coherent output text [4,5]. As such, they present an opportunity to alleviate clinicians' administrative workload by summarizing patient information contained in EHRs. While ensuring strict adherence to data privacy standards, effective models could deliver context-specific summaries that meet clinical objectives. Potential applications include optimizing information retrieval, as critical data are often buried within extensive, noisy, and repetitive entries [6,7]; or automating summarization tasks that are traditionally carried out manually, such as discharge summaries [8-10] (Figure 1). Additionally, patient-directed simplified reports could support informed decision-making [10,11].

Figure 1. The process of generating clinical summaries using LLMs. Source documents from EHRs, such as radiology reports, progress notes, nursing notes, laboratory test results, and medical reports, are processed by an LLM to generate a summary intended, in this case, for a physician. EHR: electronic health record; LLM: large language model. Created in BioRender (Bednarczyk, L., 2025, https://BioRender.com/trqp263; [12]).



Several researchers have examined the application of LLMs in health care [13-17]. Bedi et al [15], Wang et al [17], Park et al [14], and Pressman et al [16] all reported concerns regarding the strategies used to evaluate these models. Meng et al [13] highlighted the lack of robust clinical studies to validate LLMs' effectiveness and safety in real-world settings.

To the best of our knowledge, no comprehensive review has specifically addressed the performance of LLMs in clinical text summarization. This gap raises critical questions: Do LLMs genuinely outperform medical experts in summarizing clinical texts? Can they be used with confidence in clinical settings for summarization tasks?

This review seeks to assess the reliability of performance findings and their applicability to health care settings through a detailed analysis of the literature, including insights from studies such as those discussed by Van Veen et al [1].

The specific research objectives are as follows.

- Present the current state of research on clinical text summarization using LLMs.
- Evaluate the level of evidence in the current state of research.
- Assess whether these models can be used with confidence in clinical settings.
- Provide expert recommendations for current and future research.

Methods

Study Design

This scoping review focused on the summarization of clinical text within EHRs using LLMs. The authors adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist (Multimedia Appendix 1).

Search Strategy

Relevant literature, published between January 1, 2019, and June 18, 2024, was identified from 5 databases: PubMed, Embase, Web of Science, IEEE Xplore, and ACM Digital Library. The search strategy was structured around three key dimensions: "summarization," "large language models," and "healthcare," which were derived from the above-mentioned research objectives and combined with Boolean operators. Related search terms and exact database queries are presented in Multimedia Appendix 2.

Eligibility Criteria

Eligibility criteria (Textbox 1), defined prior to the screening process and agreed upon by 2 authors (DR and LB), were framed as exclusion criteria to ensure a comprehensive identification of all relevant papers. Studies that did not describe a model based on the original transformer architecture as introduced by Vaswani et al [18], did not focus on clinical text summarization, or did not engage with free-text data were excluded. We also excluded publications that were not original research, including editorials, reviews, or comments, as well as those that were not retrievable. Only peer-reviewed literature in English, French, Spanish, or German was considered. January 1, 2019, was chosen as the cutoff date based on existing literature [19].

Textbox 1. Eligibility criteria.

Exclusion criteria

- The source of evidence (SOE) does not describe a model based on the original transformer architecture.
- The SOE does not describe clinical text summarization or summarize medical texts that are not clinical (eg, biomedical texts or medical evidence summarization).
- The SOE does not deal with free-text data.
- The SOE is published before January 1, 2019.
- The SOE does not describe original research.
- The SOE is not published in English, French, Spanish, or German.
- The SOE is an editorial, review, or comment.
- The SOE is not retrievable.
- The SOE is not peer-reviewed.

Screening Process

During the initial quality check, a random sample of 10 studies, including titles and abstracts, were screened by 2 authors (LB and DR) to refine exclusion criteria and ensure consistency in screening. Discrepancies were resolved collaboratively.

The screening process was then conducted in 2 stages by the same 2 authors (LB and DR). In the first stage, titles and abstracts of the remaining studies were independently screened, with any conflicting decisions defaulting to eligibility to prevent premature exclusion of potentially relevant studies. In the second stage, full-text screening was performed independently by both authors, with any disagreements resolved through discussion and consensus.

Data Synthesis

Before data extraction, a second quality improvement phase was conducted. Two authors (LB and DR) independently extracted data from a random sample of 3 studies to refine and finalize the data extraction table. Subsequently, full-text screening and data extraction were independently carried out by three authors (LB, CGB, and AKE) using a predefined spreadsheet.

Data extraction was organized into three main aspects: (1) study context and characteristics, (2) scope of research, and (3) evaluation methodologies.

- Study context and characteristics included the year and location of publication (based on the corresponding author's address), the type of journal, the study design (inferred based on the described methodology, distinguishing between retrospective and prospective approaches), and the type of dataset used categorized as real patient data (open-source or proprietary) or synthetic data.
- The scope of research encompassed information related to the field of application, the summary intention, summarizing techniques used, technological aspects, and ethical considerations. The field of application refers to the domain where the summarization methods were developed and evaluated, covering department, country, patient demographics, and language coverage. Department and country details were extracted directly from the dataset information. Summary intention refers to the purpose of the summary, defined in this work based on the target audience, the summarization objective, and the source document. The summarization technique covered the details of input documents and the summarization techniques used. Technological aspects included relevant modeling characteristics (pretraining, fine-tuning strategies, and prompt engineering) approaches, deployment environments (eg, on-premises, cloud-based), hardware requirements, and associated computational costs. Ethical considerations included dataset deidentification and the reporting of institutional review board (IRB) approval.
- Evaluation methodologies included the strategies used, sample sizes, metrics, and additional details on each evaluation framework used.

Any discrepancies in data extraction were resolved through discussion with a fourth author (DR) to ensure accuracy and consistency throughout the process. The completed data extraction is provided in Multimedia Appendix 3 [1,21-49].

No assumption was made about missing or unclear details unless explicitly stated. The authors of the included studies were not contacted for clarification. The extraction focused exclusively on information pertinent to the summarization of clinical text found in the EHR.

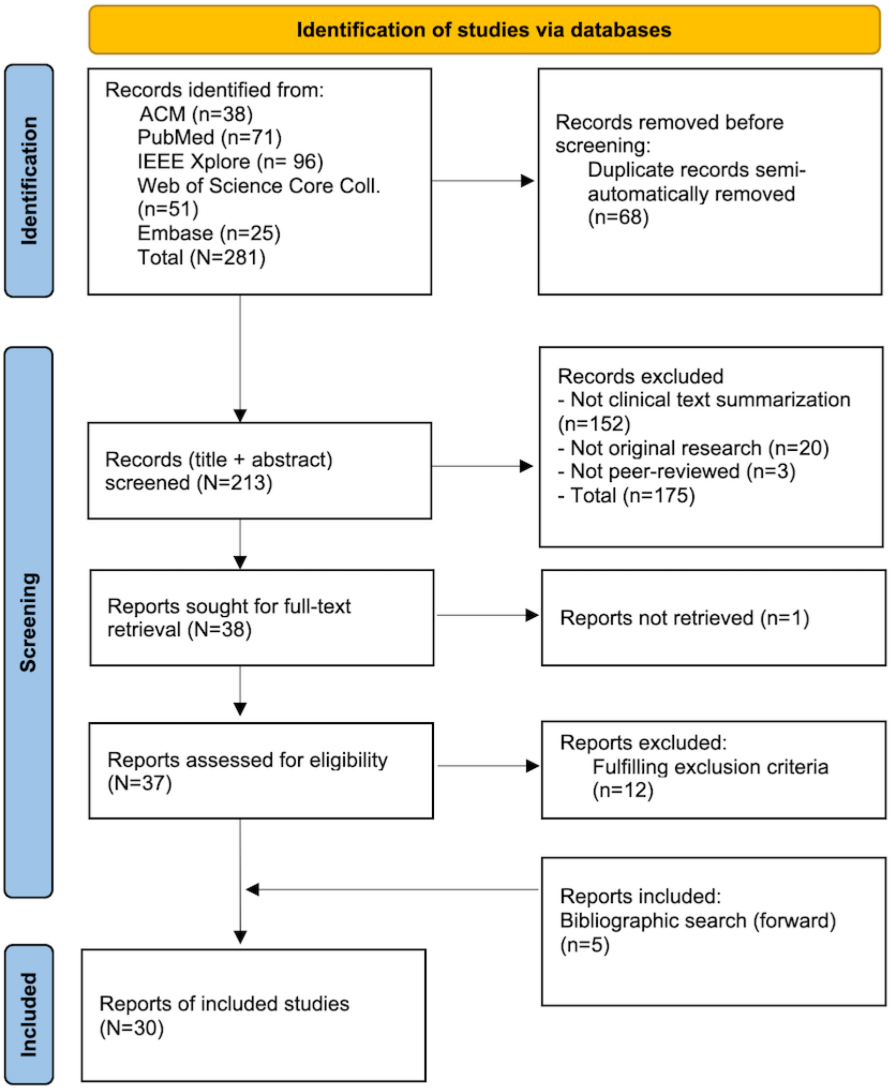
Extracted data were synthesized using a descriptive approach, complemented by narrative synthesis, and presented in tables and figures where applicable. Data were summarized and described according to the 3 categories of data extraction as described earlier. The synthesis aimed to clearly outline key trends and characteristics across the included studies without conducting statistical analyses or quantitative meta-analyses.

Results

Study Contexts and Characteristics

This scoping review included a total of 30 studies. Of the 281 retrieved by database queries, 25 were deemed eligible following title, abstract, and full-text screening. Additionally, 5 studies were incorporated through manual reference screening using the snowballing technique [20]. An overview of the literature retrieval and screening process is presented in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart (Figure 2).

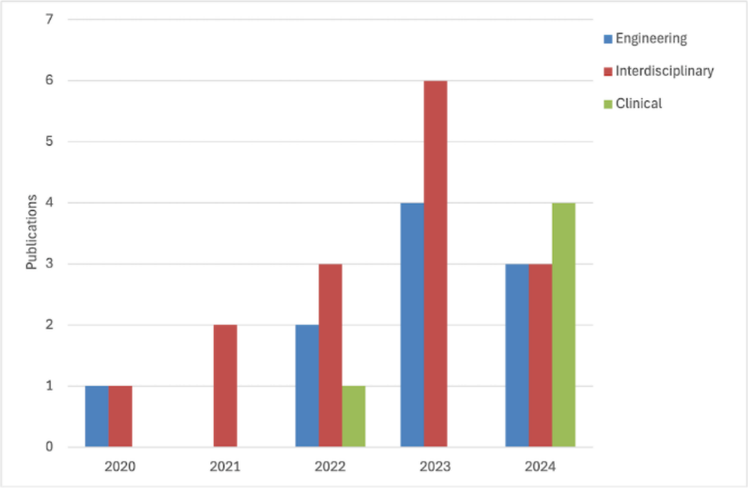
Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow chart.


Identification of studies via databases
Identification
Screening
Included
Records identified from:
ACM (n=38)
PubMed (n=71)
IEEE Xplore (n= 96)
Web of Science Core Coll. (n=51)
Embase (n=25)
Total (N=281)
Records removed before screening:
Duplicate records semi-automatically removed (n=68)
Records (title + abstract) screened (N=213)
Records excluded
- Not clinical text summarization (n=152)
- Not original research (n=20)
- Not peer-reviewed (n=3)
- Total (n=175)
Reports sought for full-text retrieval (N=38)
Reports not retrieved (n=1)
Reports assessed for eligibility (N=37)
Reports excluded:
Fulfilling exclusion criteria (n=12)
Reports included:
Bibliographic search (forward) (n=5)
Reports of included studies (N=30)

Publications were distributed across interdisciplinary (n=15, 50%), engineering (n=10, 33%), and clinical (n=5, 17%) journals. The annual publication count showed an exponential growth trend, with no study published prior to 2020 (Figure 3).

Clinical document summarization was the primary research objective in the majority of studies (n=27, 97%), while Li et al [21] used summarization as a preprocessing step in sepsis prediction.

Figure 3. Annual publication count categorized by journal type.



The study location was dominated by the United States (n=10, 33%), followed by China (n=5, 17%), India (n=3, 10%) [22-24], and the United Kingdom (n=2, 7%) [25,26]. Other locations included Australia [27], France [28], Germany [29], Lebanon [30], Spain [31], and Taiwan [32] (n=1, 3% each), and Jiang et al [33] reported multiple corresponding authors from different countries. For 10% of studies (n=3), the corresponding author's address could not be determined [34-36].

Among the studies, all used an observational retrospective design, using existing patient data to evaluate model performance. 93% used real patient data (n=28), primarily from open-source datasets (n=18, 60%), with the Medical Information Mart for Intensive Care (MIMIC) series (n=15, 50%) and the Indiana University X-Ray database (n=8, 27%) being the most reported (Table 1). Additionally, 47% of studies used proprietary databases (n=14). Goswami et al [22] mentioned using real patient data but did not specify the dataset used. Caterson et al [25] and Wu et al [28] used synthetic data, relying on scenarios generated by the authors.

Table 1. Distribution of specific open-source datasets.

Open-source dataset	Publications, n (%)
MIMIC[a]	15 (50)
IU X-Ray[b]	8 (27)
KCH[c]	1 (3)
SAM[d]	1 (3)
Stanford Coll[e]	1 (3)
PubMed	1 (3)

[a] MIMIC: Medical Information Mart for Intensive Care.

[b] IU-XRAY: Indiana University X-Ray database.

[c] KCH: King's College Hospital database.

[d] SAM: SAMSum corpus.

[e] Stanford Coll: Stanford Hospital Collection.

Scope of Research

Field of Application

Clinical departments in which the models were evaluated could be retrieved in 22 studies (73%) and included mainly the intensive care unit (ICU; n=15, 50%) neurology (n=2, 7%) [26,37], and oncology (n=2, 7%) [28,29]. Other departments included cardiology [38], geriatrics [27], neurosurgery [39], and orthopedics [25] (n=1, 3% each). Chen et al [32] and Vinod et al [24] both addressed multiple areas.

Dataset geographical origin was identified in 28 studies (93%). It was predominantly the United States (n=19, 63%) and China (n=3/30, 10%) [33,41,43]. Other origins included Germany [29], the United Kingdom [26], Taiwan [32], Spain [31], and Australia [27] (n=1, 3% each). Additionally, Searle et al [26] included patients from 2 different countries. Four studies (13%) reported on patient demographics: 3 (10%) provided information on sex, race, and ethnicity [21,39,42], and López et al [31] on the sex-age ratio of the study population.

The language coverage was mainly English (n=26, 87%), including 3 (n=3, 10%) studies explicitly stating it and 23 (n=23, 77%) inferred based on dataset sources. Additionally, 3 studies (10%) addressed the summarization of clinical documents in Chinese [33,41,43], and 1 (3%) in German [29].

Summary Intention

The intended audience of the generated summary was specified in 16 (53%) studies. Most studies targeted health care professionals (n=12, 40%), while a smaller proportion focused on patients (n=2, 7%) [30,44]. In addition, 2 studies (n=2, 7%) aimed to serve both patients and health care professionals [22,25]. Li et al [21] used summarization as a preprocessing step for subsequent modeling tasks to enhance sepsis prediction, rendering the identification of a target audience irrelevant.

Most studies (n=27, 90%) focused on a single summarization task while only a few, such as Alkhalaf et al [27], Van Veen et al [1], and Zhu et al [40], explored multiple summarization tasks. The details of summarization objectives and input sources used in each study are provided in Multimedia Appendix 4 [1,21-49].

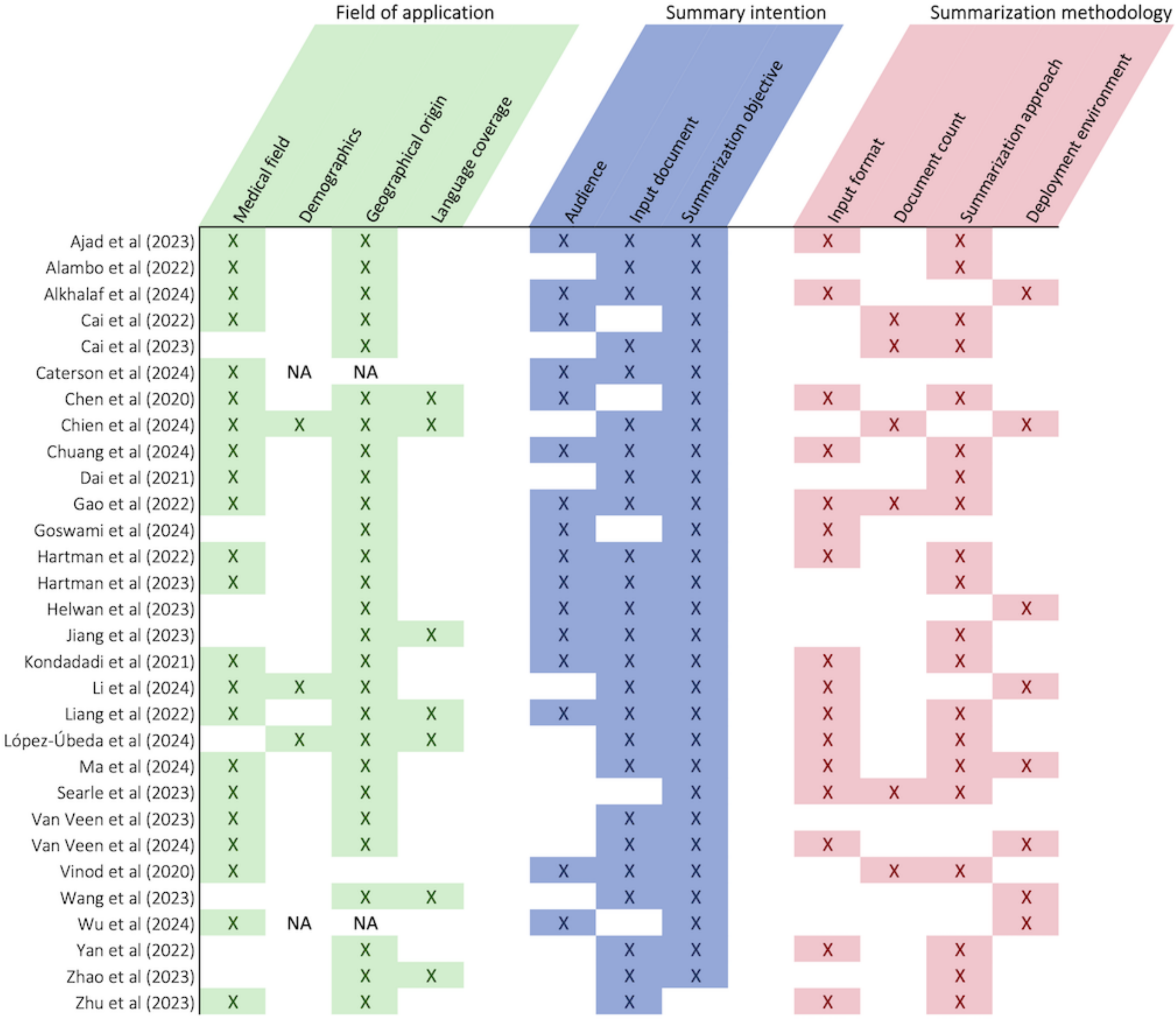
Summarization objectives included mainly generating the impression section of radiology reports (n=12, 40%), followed by generating the hospital course section of discharge summaries (n=3, 10%). Notably, the description of the summarization objectives varied across studies. Some specified the exact section of a document to be generated (eg, the impression section of radiology reports), while others described the types of source documents and the key information to be extracted [24].

Regarding input sources, most studies (n=19, 73%) used a single type of text corpus, with radiology reports being the most common (n=17, 57%), followed by progress notes (n=2, 7%) and patient forms (n=1, 3%). Overall, 23% (n=7) of studies used multiple-type text corpora. Furthermore, 13% (n=4) did not explicitly specify their source [22,26,32,44].

Summarization Methodology

Fifteen (50%) studies explicitly mentioned the input source structure: 43% (n=13) used unstructured data only, while 7% (n=2) also reported using structured data such as patient demographics [27,45]. See Figure 4 for an overview of the reported information on experimental design. Six (20%) studies explicitly reported on the number of documents used as input at once: 4 (13%) mentioned single-document summarization, Searle et al [26] specified multidocument summarization, where the model took a cluster of related documents as input, and Chien et al [39] referred to a single-multiple document approach, where multiple documents were combined and treated as a single input. 20 publications (67%) specified the summarization technique used: 17 (57%) used abstractive methods, 7 (23%) extractive methods, and 2 (7%) hybrid methods [29,37]. Among these, 5 (17%) evaluated at least 2 approaches. While abstractive summarization generates new sentences that paraphrase the core ideas of the source text, extractive summarization selects and compiles existing key sentences or phrases directly from the original content [50].

Figure 4. Overview of the reported information, including the medical field, dataset demographics, geographical origin of the test set, language coverage of the study, the intended audience of the summary (eg, physician, patient), a document used as input, summarization objective, input format (structured, unstructured, and both), input document count (single, multiple, and single-multiple), summarization approach (abstractive, extractive, and hybrid), and deployment environment (on-premises and cloud-based) [1,21-49]. NA: not available.



	Field of application				Summary intention			Summarization methodology			
	Medical field	Demographics	Geographical origin	Language coverage	Audience	Input document	Summarization objective	Input format	Document count	Summarization approach	Deployment environment
Ajad et al (2023)	X		X		X	X	X	X		X	
Alambo et al (2022)	X		X			X	X			X	
Alkhalaf et al (2024)	X		X		X	X	X	X			X
Cai et al (2022)	X		X		X		X		X	X	
Cai et al (2023)			X			X	X		X	X	
Caterson et al (2024)	X	NA	NA		X	X	X				
Chen et al (2020)	X		X	X	X		X	X		X	
Chien et al (2024)	X	X	X	X		X	X		X		X
Chuang et al (2024)	X		X		X	X	X	X		X	
Dai et al (2021)	X		X			X	X			X	
Gao et al (2022)	X		X		X	X	X	X	X	X	
Goswami et al (2024)			X		X		X	X			
Hartman et al (2022)	X		X		X	X	X	X		X	
Hartman et al (2023)	X		X		X	X	X			X	
Helwan et al (2023)			X		X	X	X				X
Jiang et al (2023)			X	X	X	X	X			X	
Kondadadi et al (2021)	X		X		X	X	X	X		X	
Li et al (2024)	X	X	X			X	X	X			X
Liang et al (2022)	X		X	X	X	X	X	X		X	
López-Úbeda et al (2024)		X	X	X		X	X	X		X	
Ma et al (2024)	X		X			X	X	X		X	X
Searle et al (2023)	X		X				X	X	X	X	
Van Veen et al (2023)	X		X			X	X				
Van Veen et al (2024)	X		X			X	X	X			X
Vinod et al (2020)	X				X	X	X		X	X	
Wang et al (2023)			X	X		X	X				X
Wu et al (2024)	X	NA	NA		X		X				X
Yan et al (2022)			X			X	X	X		X	
Zhao et al (2023)			X	X		X	X			X	
Zhu et al (2023)	X		X			X		X		X	

Twenty-six (87%) studies reported on open-source models, with bidirectional encoder representations from transformers–based models being the most frequently cited (n=12, 40%). Seven (23%) studies investigated proprietary models, all based on commercial services provided by OpenAI. Two (7%) examined both model types [1,39], while Li et al [21] used an ensemble model. Seven (23%) studies performed additional pretraining of existing model architectures prior to fine-tuning. Furthermore, 23 (77%) mentioned fine-tuning. Ten (33%) studies conducted prompt engineering.

Eight (27%) studies reported on the deployment environment: 3 (10%) indicated on-premise deployment [27,30,39], and 6 (20%) the use of external services, including Azure OpenAI application programming interface (API) [1], and OpenAI API [21,28,36,39,41]. Sixteen (53%) studies reported on hardware requirements, whereas none mentioned associated costs.

Ethical Considerations

Deidentification was conducted in 13 (43%) studies, and anonymization in 2 (7%) studies. Additionally, Wang et al [41] mentioned "data desensitization," and López et al [31] mentioned ensuring that no patient or doctor information was included in datasets. IRB approval was reported in 30% (n=9) of studies.

Evaluation Methodology

Evaluation Approaches

External validation, which involves testing models on datasets not used during training, was not explicitly reported in most cases. However, 2 (7%) studies reported testing on unseen datasets during model development (training or fine-tuning) and relied solely on automated metrics for external evaluation [35,46]. Internal validation, which assesses a model's performance using a dedicated sample of the dataset it was trained on, was conducted in all studies.

Global performance evaluation, defined as the evaluation process conducted on the full test set, was conducted was conducted all studies. Furthermore, 16 (53%) studies used both automatic validation metrics and human evaluations, while 10 (33%) used automatic metrics only, and 4 (13%) human evaluations only. Subgroup performance evaluation was reported in 2 (7%) papers. Out-of-distribution performance testing, a process that evaluates how well a model performs on data that are underrepresented in the training set [51], was reported by Van Veen et al [47]. Subcategory testing was reported by Liang et al [29], who evaluated the impact of different stages of cancer on model performance using patient degree matching.

Model failure analysis was conducted through error categorization in 6 (20%) studies. Four (13%) studies addressed both the classification and quantification of error and 2 (7%) focused exclusively on error classification [44,47]. Bias analysis remained unexplored, as no studies explicitly report structured bias assessments. Finally, patient safety risk analysis was assessed by Van Veen et al [1] (n=1, 3%) using a Healthcare Adapted Risk Management Scale [1].

A broader understanding of model performance was achieved through various approaches, including ablation studies (n=8, 27%) and attention distribution analysis (n=2, 7%) [33,34]. Additionally, Zhao et al [43] included analyses of sparsity, loss curves, and the Zipf distribution, while Li et al [21] evaluated model performance through a proxy task involving sepsis prediction.

Automated Performance Metrics

Automated performance metrics were reported in 26 (87%) papers. Test set size, mentioned in 23 (77%) studies, included mainly between 15 and 1000 documents (0-1000 documents: 33%; 1000–5000 documents: 30%; more than 5000 documents: 10%). Randomized sampling was explicitly mentioned in 5 (17%) publications, while 3 used the entire dataset as a test set (10%) [25,39,42].

Reference-based metrics were widely used (n=25, 77%). Specifically, this type of metrics compared generated summaries to predefined reference summaries that encompassed the original document (n=18, 60%), manually labeled data (n=4, 13%), and expert-generated summaries (n=2, 7%) [1,39]. Two (7%) studies did not mention the reference summary used [22,26]. Validation of reference summaries was conducted in 7 (23%) studies, including 1 (3%) using the original document as a reference summary [43]. Recall-Oriented Understudy for Gisting Evaluation score was the most frequent performance measure (n=24, 80%), followed by Bilingual Evaluation Understudy (n=8, 27%) and bidirectional encoder representations from transformers score (n=7, 23%).

The test set count and evaluation metrics for both automatic metrics and human evaluation in each publication are detailed in Multimedia Appendix 5 [1,21-49].

Human Evaluation

Human assessment was reported in 20 (67%) papers. Test set size, mentioned in 18 (60%) studies, included mainly between 2 and 50 documents (2-50 documents: n=10, 33%; 50–100 documents: n=6, 20%; more than 100 documents: n=2, 7%). Randomized sampling was explicitly mentioned in 12 (40%) publications, including Li et al [21] who specified the use of 5-fold cross-validation. Additionally, 3 (10%) studies used the entire dataset [25,39,41].

Metrics, detailed in 18 (60%) studies, encompassed readability (n=12, 40%), factual correctness (n=12, 40%), and the adequacy of provided information directly related to the summary intention (n=12, 40%). These assessments include relevance [26,31], completeness [1,37] (n=2, 7% each), the ability to capture critical information [47], adequacy [44], informativeness [34], omission or insertion [25], comprehensiveness [39], and effectiveness [24] (n=1, 3% each). Additionally, clinical use was assessed in 3 (10%) studies and involved estimated clinical time saved [28], ease of revision [44], and clinical use [43].

Blind analysis was reported in 6 (20%) publications. The assessor recruitment process was documented by Jiang et al [33] and Li et al [21] (n=2, 7%), who specified including volunteers and invitees respectively. Assessors' affiliation, reported in 16 (53%) studies varied. Most studies (n=14, 47%) included at least 1 in-domain physician. The inclusion of several participants in the manual review process was common (n=14, 47%), with 9 (30%) studies specifically involving at least 2 in-domain physicians.

Ten studies (33%, 10/30) reported at least 2 raters per document. Interannotator agreement score was reported in 4 (13%) studies and included intraclass correlation [1,37], Pearson *r* of human evaluation scores [44], and Cohen κ [26]. The use of measurement scales, documented in 14 papers (47%), involved mainly numeric scores (n=7, 23%) and Likert Scale (n=6, 20%). Additionally, Cai et al [44] used a scoring scale [44].

Three (10%) studies mentioned following a specific protocol for human evaluation: Chien et al [39] applied the approach proposed by Goldstein et al [52], and Lopez et al [31] and Searle et al [26] followed the protocol SummEval proposed by Fabbri et al [53]. The study setup was documented in 6 (20%) studies, where authors described the tools used in the research process, such as interfaces and questionnaires.

Summary ranking methods were primarily independent (n=19, 63%), where items are evaluated independently without direct comparison to one another. In contrast, Lopez et al [31] used a pairwise comparison approach, where items are evaluated by directly comparing 2 options at a time, allowing for a more relative assessment of rankings.

Discussion

Principal Findings

This scoping review offers a comprehensive overview of the state of research on clinical text summarization using LLMs. Concerns have been raised about the applicability of certain machine learning models in clinical settings, where their effectiveness failed to meet real-world expectations [54-56]. Our analysis suggests that similar challenges may extend to the field of automated summarization. The following sections discuss findings, highlight key barriers in translating research findings into practical clinical applications, and propose directions for future research.

A Narrow Research Scope

Current research focuses on a limited range of summarization objectives, patient populations, and medical specialties. Most studies (n=17, 57%) focused on radiology reports, particularly the impression section (n=15, 50%). Study populations were predominantly ICU patients (n=15, 50%), from US-based institutions (n=19, 73%). Language coverage was predominantly English (n=26, 87%). MIMIC, an open-source dataset originating from ICU from Beth Israel Deaconess Medical Center was used in 50% of studies (Figure 4) [57].

The limited research scope, combined with the heavy reliance on a few publicly available datasets, raises concerns about whether research is driven by clinical needs or data availability. ICU patients constitute only a small subset of the broader health care population encountered in routine clinical practice, limiting the generalizability of findings. The practical necessity of automating the impression section in radiology reports is debatable [58].

Additionally, open-source datasets fail to capture real-world complexities such as variations in writing styles, clinical workflows, and patient populations [59,60]. As a result, many potential applications remain yet unexplored. Future research should ensure alignment with clinical needs, and expand its scope to include a wider range of use cases, medical specialties, patient populations, and language coverage (Textbox 2).

Textbox 2. Key research priorities in study elaboration for large language model (LLM)–based summarization. This textbox highlights critical areas for refining and expanding research methodologies to improve clinical validity, ethical and legal compliance, and practical deployment of summarization models in health care.

Research objectives

- Ensure research objectives directly address real-world clinical challenges.
- Expand the scope of investigation by diversifying use cases, medical specialties, and patient populations to enhance practical applicability.
- Clearly define the context of summarization objectives, including purpose, target audience, and expected outcomes, to ensure a clinically valid evaluation.

Summarization methodology

- Expand knowledge of both the performance and limitations of LLMs by evaluating their ability to handle real-world complexities, such as multisource summarization.

Model performance and clinical impact evaluation

- Conduct a context-aware validation of the model (or task-specific evaluation) through human assessment.
- Assess model generalizability via external validation, subcategory testing, and k-fold cross-validation.
- Evaluate model utility by analyzing its impact on the intended audience to determine whether summarization systems effectively fulfill their clinical purpose.
- Perform failure analysis and subsequent patient safety risk assessment to evaluate the clinical impact of identified errors on patient safety.
- Conduct bias analysis by assessing outputs for discriminative language, including bias related to gender, race, socioeconomic status, substance use, and mental health, to ensure fairness.

Address data security and privacy

- Ensure compliance with legal agreements, and implement data protection strategies such as on-premises deployment to safeguard patient privacy and regulatory standards.

Another limitation observed in this research is the lack of a clear description of summarization objectives. For example, Zhu et al [40] described objectives as “to generate a few significant impressions” or “to generate a few critical diagnosis results” (Multimedia Appendix 4 [1,21-49]). These descriptions lack precision, making it difficult to assess what constitutes a “significant” or “critical” diagnosis result, and how the output should be evaluated. As such, comparing studies and assessing model effectiveness becomes difficult.

Poorly defined research objectives can also lead to irrelevant research or fail to ensure proper model validation. For example, Helwan et al [30] identified patients as the target audience and summarizing and simplifying radiology reports as the summarization objective. In contrast, their application focused on generating the impression section of radiology reports primarily meant for communication between the ordering physician and the radiologist [61]. This misalignment between the stated objective and the actual application further highlights the need for greater clarity in research goals.

Future research should define summarization objectives, specifying purpose, target audience, and expected outcomes, to ensure a clinically relevant evaluation by aligning assessment criteria with intended clinical applications (Textbox 2). Additionally, studies should provide a detailed text corpus description, ensuring clear and consistent terminology across regions, disciplines, and institutions to facilitate comparability across studies (Textbox 3).

Textbox 3. Reporting recommendations specific to the large language model (LLM)–based summarization research.

Research objective

- Provide a thorough description of the summarization objective studied, a precise understanding of its purpose, intended audience, and expected outcome.

Summarization methodology and model evaluation

- Provide a clear and detailed description of the text corpus used, ensuring that terminology is well-defined and understandable across different regions, disciplines, and institutions.
- Specify the structure of the input test data (structured and unstructured).
- Report the number of documents processed by the model (single, multiple, and multiple-single).
- Detail the summarization technique used (abstractive, extractive, and hybrid).
- Specify the model deployment environment (on-premises and externally).
- Provide demographic information on the dataset used for LLM development, tuning, or evaluation.

Technological and cost requirements

- Provide a detailed report on the computational resources needed for model deployment and outline the associated costs.

Limitations in Addressing Real-World Effectiveness

Most studies used abstractive summarization (n=17, 57%), typically with single-document inputs (n=4, 13%) and unstructured data (n=13, 43%). In terms of deployment methods, 10% (n=3) of models deployed on-premises and 20% (n=6) used external cloud services. Furthermore, as depicted in Figure 4, notable underreporting across studies was noted.

These findings highlight concerns about model effectiveness, particularly in synthesizing multisource data and handling longitudinal records. In clinical practice, physicians summarize patient data from single or multiple sources using different perspectives for decision-making, communication, or documentation. However, only Searle et al [26] specified multidocument summarization, while Chien et al [39] mentioned a single-multiple document approach. Multidocument summarization presents additional challenges, such as maintaining coherence, reducing redundancy, and ensuring consistency [62,63]. Furthermore, temporality remains a key obstacle, as clinical narratives span longitudinal records. Chien et al [39] suggested that overcoming this challenge may require dedicated models for temporal relation extraction to accurately capture event sequencing. Thus, the full spectrum of summarization performance remains largely unexplored. Future research should deepen insights into LLM strengths and weaknesses by assessing their ability to manage real-world challenges, including multisource summarization (Textbox 2).

Additionally, concerns persist regarding the real-world applicability of current summarization models due to deployment constraints. On-premise deployment presents hardware limitations that must be addressed to maintain optimal performance and accuracy, whereas cloud-based models necessitate stringent data privacy measures and compliance with regulatory frameworks [64,65]. Without systematic reporting on these deployment considerations, it remains uncertain whether summarization models can be practically integrated into health care workflows. Identifying practical strategies for real-world implementation is essential to bridge the gap between research and clinical application.

Finally, the widespread under-reporting across studies undermines the validity of performance assessments (Figure 4), making it difficult to compare models, reproduce results, and evaluate their real-world feasibility. To address this, future research must adopt comprehensive reporting guidelines, such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD)+LLM checklist, which extends the original TRIPOD guidelines to ensure standardized reporting in LLM-based health care research [66]. Additionally, Textbox 3 outlines reporting recommendations specific to summarization research. Enhancing transparency in reporting will improve integrity, enable meaningful cross-study comparisons, and ultimately support the development of summarization models that are both clinically relevant and practically deployable.

Challenges and Limitations in Model Evaluation

Evaluation frameworks varied widely in used strategies and metrics (Multimedia Appendix 5 [1,21-49]). All studies conducted internal validation, primarily with automatic metrics (n=26, 87%), while 67% (n=20) included human evaluation. Few studies conducted external validation (n=2, 7%), failure analysis (n=6, 20%), or patient safety risk assessments (n=1, 3%), and none performed bias evaluation.

Challenges in Ensuring Reliable Evaluations

Effectively evaluating generative outputs in summarization tasks remains a challenge, as it requires evaluating both textual quality and contextual appropriateness [66]. Human evaluation remains the gold standard for text summarization, offering the contextual insight necessary for accurate summary assessment [67-69]. Automatic validation metrics, however, lack contextual understanding and therefore cannot reliably determine relevance, correlate poorly with human judgment, and are considered underinformative [67,70].

Findings suggest that a notable portion of studies (33%) may lack a clinically meaningful assessment of model performance, as they relied solely on automatic validation metrics. Several studies reported limitations of automatic validation metrics in summary evaluation. Cai et al [44] and Hartmann et al [45] noted that these metrics were not adequate to capture summary quality, while Liang et al [29] and Zhao et al [43] reported that these metrics did not evaluate the clinical validity (or usability) of summaries. As no consensus is reached on the reliability of automatic validation metrics in evaluating abstractive summarization, several studies have proposed combining automatic metrics with human evaluation and conducting correlation analyses to better assess their validity and practical value [66].

Further limitations were observed among the 20 studies conducting human evaluations: 15% (3/20) did not report the number of assessors [22,27,43], 20% (4/20) did not specify the assessor's affiliation [22,27,41,43], 90% (18/20) did not specify their assessors' recruitment processes, and 70% (14/20) of studies lacked blinded analysis. While 50% (10/20) of studies reported dual annotations, only 4 out of 20 (20%) reported interannotator agreement. Addressing these issues is essential to ensure reliable performance assessment. While yet no standardized best practices for human evaluation in text summarization currently exist, research in this area is progressing in a promising direction. Tam et al [71] proposed an evaluation framework aimed at enhancing reliability, generalizability, and applicability in human evaluation practices. Similarly, Van der Lee et al [67] introduced a set of best practices for the manual review of the automatically generated text, contributing to the development of more structured and consistent evaluation methodologies.

Challenges in Evaluating the Robustness of Model Performance

Assessing model robustness is crucial, as clinical settings differ in patient populations, writing styles, and medical practices. However, several issues were identified in the studies reviewed, particularly regarding the depth and comprehensiveness of the evaluation processes.

The sample sizes used in human evaluation were limited. Most studies assessed model performance on as few as 2 to 50 documents. Since LLMs generate nondeterministic outputs, they can vary between each iteration [72]. A previous work by Tam et al [71] suggests that at least 130 documents should be evaluated when testing clinical decision support tools to enable meaningful performance assessments.

In addition, model generalizability, referring to the model's ability to perform effectively across diverse clinical settings, populations, or conditions beyond those on which it was originally trained [55], was frequently overlooked in the analyzed studies. External validation, essential for detecting overfitting and assessing cross-site transportability [73], was only reported by Kondadadi et al [35] and Dai et al [46]. Similarly, subgroup performance analysis was conducted in another 2 studies: Liang et al [29] evaluated cancer stage-specific performance using subpopulation testing, and Van Veen et al [47] conducted out-of-distribution testing, which involves evaluating the model data on underrepresented samples of the dataset [51]. Additionally, 5-fold cross-validation, which involves partitioning the training data into different subsets to ensure consistent model performance, was only used by Li et al [21].

Assessing model generalizability involves understanding how data variability affects model performance [74]. Distribution shifts in training data can lead to underperformance in underrepresented populations while causing overfitting in overrepresented subgroups [51]. As such, while models may demonstrate high performance within their training environments, their ability to generalize to broader clinical settings remains uncertain.

To ensure robustness, future research should prioritize generalizability assessments, including rigorous out-of-distribution testing and subcategory analysis within the same dataset, even when access to multiple external datasets is limited. Moreover, systematic reporting of demographic characteristics is essential for enhancing transparency and ensuring models are evaluated across diverse populations (Textbox 2) [54,66,75,76].

Limitations in Addressing the Clinical Impact of Summarization Models

Clinical Utility

Few studies assessed the clinical impact of LLM-based summarization. Wu et al [28] evaluated perceived benefits by clinicians, focusing on estimated time savings. Cai et al [44] examined the ease of revision, while Zhao et al [43] assessed the model's clinical utility.

LLMs are expected to reduce clinician workload and improve information synthesis, raising high adoption expectations [75]. However, beyond assessing their technical performance, a thorough evaluation of both their utility and risks is crucial to inform responsible implementation and compliance with legal standards [77,78]. Future studies should determine whether summarization systems effectively serve their intended clinical purpose [77]. Additionally, since all studies included in this review were retrospective, prospective studies could provide a deeper understanding of the actual impact of LLM-based summarization models in clinical workflows [56].

Patient Safety Risks

The safety and potential clinical harm of evaluated models remained frequently unaddressed. Failure analysis was conducted in only 20% (n=6) of studies, with Van Veen et al [1] being the only study to examine patient safety risks.

Without a clear understanding of failure patterns, it becomes inherently difficult to assess their impact on clinical decision-making and patient safety, as does the development of effective safeguards to mitigate potential risks. While recent studies have attempted to establish taxonomies for hallucinations [79,80], failures may be inherently task-specific, underscoring the need for systematic error analysis and categorization based on existing research. Additionally, their impact on patient safety should be evaluated using appropriate risk analysis methods (Textbox 2).

Although no standardized risk assessment framework currently exists, a step toward addressing this gap is the risk matrix-based evaluation framework introduced by Asgari et al [81], which provides a structured methodology for categorizing and quantifying errors, enabling a systematic assessment of their impact on patient safety. Establishing standardized methodologies for failure analysis and risk assessment will be essential to ensure the safe and effective deployment of these models in clinical practice.

Bias and Fairness

LLMs risk exacerbating health disparities [82] as they absorb intrinsic biases during training on diverse data sources. These biases can manifest as harms in specific downstream tasks, impacting clinical decision-making [5,10,54,60,76,83]. For example, Zack et al [84] identified stereotypical demographic representations in LLM-generated diagnostic and treatment recommendations related to sex, ethnicity, and race. As a generative process, abstractive summarization also raises the potential to perpetuate these biases [83].

Despite these concerns, no study reported a biased assessment. Future research should use diverse datasets, prioritize bias detection, particularly in identifying discriminative language in model outputs, and develop mitigation strategies to ensure fairness in clinical decision support (Textbox 2) [54,75,76].

Data Privacy and Security

Ensuring data privacy and security is essential for ethical and legal compliance in research, and to enable valid and reproducible studies that can inform the responsible adoption of LLMs for clinical summarization. However, several studies used proprietary models to process patient datasets, raising concerns about data privacy and regulatory compliance.

For example, Van Veen et al [1], Ma et al [36], and Li et al [21] used the MIMIC series, however, only Van Veen et al [1] explicitly reported using the Azure OpenAI API. This raises concerns as the PhysioNet Credentialed Data Use Agreement explicitly prohibits sharing credentialed datasets (eg, MIMIC-III, MIMIC-IV, MIMIC-CXR) with third-party AI services, such as OpenAI APIs [85]. Chien et al [39] and Wang et al [41] used deidentified proprietary datasets. However, Wang et al [41] used these datasets without IRB approval, raising ethical and regulatory concerns regarding research oversight, patient privacy, and compliance with data protection standards.

Model selection must balance multiple factors, including performance, regulatory compliance, ethical considerations, and socioeconomic factors. Proprietary models, such as GPT-4, are attractive as they offer strong performance, ease of use, and cost-efficiency. However, there are currently there are currently no clear regulatory guidelines or ethical consensus on handling patient information in private models [5,86].

While Ma et al [36] argue that deidentification sufficiently protects patient privacy, studies have demonstrated that deidentified data could be reidentified, raising concerns about its reliability as a sole privacy safeguard [87]. Despite ongoing advancements in data protection methods, vulnerabilities persist, with emerging privacy attacks continuously exposing weaknesses in data protection measures [88].

Given these challenges, future research must explicitly disclose data-sharing practices, ensure compliance with legal agreements, and adopt privacy-preserving strategies, such as on-premises deployment (Textbox 2). Locally installed open-source solutions provide a controlled environment tailored to institutional needs, ensuring strict data privacy compliance while supporting valid, reproducible studies that can lead to practical solutions for reducing clinicians' workload.

Technological and Cost Requirements

The successful deployment of LLMs also depends on their computational resource requirements and cost implications. While most studies reported hardware and memory requirements, none provided a detailed cost analysis.

Without these insights, health care institutions may struggle to assess the feasibility of implementing LLM-based solutions. Future research should go beyond hardware specifications to evaluate the economic impact of deployment to ensure that LLM adoption is both technically and financially sustainable in clinical settings (Textbox 3).

Strengths and Limitations

The data extraction process was conducted independently by 3 authors, which enhances the accuracy and reliability of the results. By aligning with the PRISMA-ScR checklist, we ensure transparency throughout the review, allowing for a coherent and well-documented process.

However, several limitations must be acknowledged. First, to finalize the data extraction table, we randomly assessed 3 studies, following the methodology outlined by Pollock et al [89]. This may have introduced bias in defining extracted variables. Second, certain aspects of data extraction were not fully explored in this study, such as the architecture types of models and their performance. This decision was made to maintain a broader focus rather than a highly technical approach. Additionally, providing a detailed performance evaluation was not meaningful due to the limitations mentioned in the discussion. Third, as authors were not contacted, reliance on published information alone may have introduced bias, particularly where methodological details were incomplete or underreported. Finally, the rapid pace of new research may result in some emerging studies being missed, and relevant sources may be inaccessible due to publication bias or their status as preprints.

Conclusions

This scoping review highlights key barriers to translating research advancements into practical applications, indicating that the field is still in its early stages. Research remains limited in scope, often shaped by dataset availability rather than explicitly guided by clinical needs, leaving many potential applications unexplored. Performance assessments frequently lack reliability and robustness, making it difficult to accurately evaluate model effectiveness. Furthermore, clinical impact evaluations remain insufficient, raising concerns about model utility, potential risks, fairness, data privacy, and broader technological and cost implications.

To advance this field, future research must broaden its scope, strengthen methodological transparency, and improve the reliability of evaluation frameworks. Additionally, enhancing model robustness and conducting comprehensive clinical impact assessments will be essential for determining the practical value of LLM-based summarization.

Authors' Contributions

LB conceptualized the study; developed the methodology; extracted and analyzed the data, wrote the first draft, and reviewed and edited the manuscript. DR conceptualized the study; developed the methodology; wrote the first draft and reviewed the manuscript. CGB conceptualized the study; extracted and analyzed the data; and reviewed the manuscript. AKE extracted and analyzed the data; wrote the first draft and reviewed the manuscript. YZ, AB, and MB conceptualized the study and reviewed the manuscript. CL supervised the review. All authors have full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. All authors gave final approval.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 404 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search queries from databases (as of June 18, 2024).
[DOCX File , 15 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Data extraction table.
[XLSX File (Microsoft Excel File), 190 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Summary of input sources and corresponding summarization objectives for each publication.
[DOCX File , 53 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Additional information on results: test set size and specific evaluation metrics reported for each publications.
[DOCX File , 24 KB-Multimedia Appendix 5]

References

1. Van Veen D, Van Uden C, Blankemeier L, Delbrouck J, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
2. Keszthelyi D, Gaudet-Blavignac C, Bjelogrlic M, Lovis C. Patient information summarization in clinical settings: scoping review. JMIR Med Inform. 2023;11:e44639. [FREE Full text] [doi: 10.2196/44639] [Medline: 38015588]
3. Schnipper J, Fitall E, Hall K, Gale B. Approach to improving patient safety: Communication. PSNet. 2020. URL: https://psnet.ahrq.gov/perspective/approach-improving-patient-safety-communication [accessed 2024-09-10]

4. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. Nat Rev Phys. 2023;5(5):277-280. [doi: 10.1038/s42254-023-00581-4]
5. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. Jul 06, 2023;6(1):120. [FREE Full text] [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
6. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC Bioinformatics. Jan 16, 2013;14:10. [FREE Full text] [doi: 10.1186/1471-2105-14-10] [Medline: 23323800]
7. Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. J Biomed Inform. Aug 2011;44(4):688-699. [FREE Full text] [doi: 10.1016/j.jbi.2011.03.008] [Medline: 21440086]
8. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-180. [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
10. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med. 2023;3(1):141. [FREE Full text] [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
11. Marinello D, Di Cianni F, Del Bianco A, Mattioli I, Sota J, Cantarini L, et al. Empowering patients in the therapeutic decision-making process: a glance into Behçet's syndrome. Front Med. 2021;8:769870. [FREE Full text] [doi: 10.3389/fmed.2021.769870] [Medline: 34966756]
12. Bednarczyk L. Figure 1. BioRender. 2025. URL: https://app.biorender.com/citation/68078b899c45b69d2a8640db [accessed 2024-04-22]
13. 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. [FREE Full text] [doi: 10.1016/j.isci.2024.109713] [Medline: 38746668]
14. Park Y, 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. [FREE Full text] [doi: 10.1186/s12911-024-02459-6] [Medline: 38475802]
15. 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):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
16. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider CR, Forte AJ. Clinical and surgical applications of large language models: a systematic review. J Clin Med. 2024;13(11):3041. [FREE Full text] [doi: 10.3390/jcm13113041] [Medline: 38892752]
17. Wang L, Wan Z, Ni C, Song Q, Li Y, Clayton E, et al. A systematic review of ChatGPT and other conversational large language models in healthcare. MedRxiv. Preprint posted online on April 27, 2024. 2024. [FREE Full text] [doi: 10.1101/2024.04.26.24306390] [Medline: 38712148]
18. Vaswani A, Shazeer N, Parmar N. Attention is All you Need. Advances in Neural Information Processing Systems. United Kingdom. Curran Associates, Inc; 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [accessed 2024-07-23]
19. Zhang H, Yu PS, Zhang J. A systematic survey of text summarization: from statistical methods to large language models. ACM Comput Surv. 2025. [doi: 10.1145/3731445]
20. Wohlin C, Kalinowski M, Romero Felizardo K, Mendes E. Successful combination of database search and snowballing for identification of primary studies in systematic literature studies. Inf Softw Technol. 2022;147:106908. [doi: 10.1016/j.infsof.2022.106908]
21. Li Q, Ma H, Song D, Bai Y, Zhao L, Xie K. Early prediction of sepsis using chatGPT-generated summaries and structured data. Multimed Tools Appl. 2024;83(41):89521-89543. [doi: 10.1007/s11042-024-18378-7]
22. Goswami J, Prajapati KK, Saha A, Saha AK. Parameter-efficient fine-tuning large language model approach for hospital discharge paper summarization. Appl Soft Comput. 2024;157:111531. [doi: 10.1016/j.asoc.2024.111531]
23. Ajad A, Saini T, Niranjan K. Rad-formertructuring radiology reports using transformers*. 2023. Presented at: 5th International Conference on Recent Advances in Information Technology (RAIT); March 3, 2023:1-6; Dhanbad, India.
24. Vinod P, Safar S, Mathew D, Venugopal P, Joly L, George J. Fine-tuning the BERTSUMEXT model for clinical report summarization. 2020. Presented at: International Conference for Emerging Technology (INCET); June 5-7, 2020; Belgaum, India.
25. Caterson J, Ambler O, Cereceda-Monteoliva N, Horner M, Jones A, Poacher AT. Application of generative language models to orthopaedic practice. BMJ Open. 2024;14(3):e076484. [FREE Full text] [doi: 10.1136/bmjopen-2023-076484] [Medline: 38485486]
26. Searle T, Ibrahim Z, Teo J, Dobson RJ. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. J Biomed Inform. 2023;141:104358. [FREE Full text] [doi: 10.1016/j.jbi.2023.104358] [Medline: 37023846]

27. Alkhalaf M, Yu P, Yin M, Deng C. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. J Biomed Inform. 2024;156:104662. [FREE Full text] [doi: 10.1016/j.jbi.2024.104662] [Medline: 38880236]
28. Wu DJ, Bibault J. Pilot applications of GPT-4 in radiation oncology: Summarizing patient symptom intake and targeted chatbot applications. Radiother Oncol. 2024;190:109978. [doi: 10.1016/j.radonc.2023.109978] [Medline: 37913954]
29. Liang S, Kades K, Fink M. Fine-tuning BERT models for summarizing German radiology findings. 2022. Presented at: Proceedings of the 4th Clinical Natural Language Processing Workshop; July 1, 2022:30-40; Seattle, WA.
30. Helwan A, Azar D, Ozsahin D. Medical reports summarization using text-to-text transformer. 2023. Presented at: Advances in Science and Engineering Technology International Conferences (ASET); February 20-23, 2023; Dubai, UAE.
31. López-Úbeda P, Martín-Noguerol T, Díaz-Angulo C, Luna A. Evaluation of large language models performance against humans for summarizing MRI knee radiology reports: a feasibility study. Int J Med Inform. 2024;187:105443. [doi: 10.1016/j.ijmedinf.2024.105443] [Medline: 38615509]
32. Chen Y, Chen Y, Lin J, Huang C, Lai F. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR Med Inform. 2020;8(4):e17787. [FREE Full text] [doi: 10.2196/17787] [Medline: 32347806]
33. Jiang Z, Cai X, Yang L, Gao D, Zhao W, Han J, et al. Learning to summarize Chinese radiology findings with a pre-trained encoder. IEEE Trans Biomed Eng. 2023;70(12):3277-3287. [doi: 10.1109/TBME.2023.3280987] [Medline: 37314905]
34. Cai X, Liu S, Han J, Yang L, Liu Z, Liu T. ChestXRayBERT: A pretrained language model for chest radiology report summarization. IEEE Trans Multimedia. 2023;25:845-855. [doi: 10.1109/tmm.2021.3132724]
35. Kondadadi R, Manchanda S, Ngo J, McCormack R. Optum at MEDIQA 2021: Abstractive summarization of radiology reports using simple BART finetuning. 2021. Presented at: Proceedings of the 20th Workshop on Biomedical Language Processing; June 1, 2021; Virtual meeting.
36. Ma C, Wu Z, Wang J, Xu S, Wei Y, Liu Z, et al. An iterative optimizing framework for radiology report summarization with ChatGPT. IEEE Trans Artif Intell. 2024;99:1-12. [doi: 10.1109/tai.2024.3364586]
37. Hartman VC, Bapat SS, Weiner MG, Navi BB, Sholle ET, Campion TR. A method to automate the discharge summary hospital course for neurology patients. J Am Med Inform Assoc. 2023;30(12):1995-2003. [doi: 10.1093/jamia/ocad177] [Medline: 37639624]
38. Alambo A, Banerjee T, Thirunarayan K, Cajita M. Improving the factual accuracy of abstractive clinical text summarization using multi-objective optimization. Annu Int Conf IEEE Eng Med Biol Soc. 2022;2022:1615-1618. [doi: 10.1109/EMBC48229.2022.9871798] [Medline: 36085755]
39. Chien A, Tang H, Jagessar B, Chang K, Peng N, Nael K, et al. AI-assisted summarization of radiologic reports: evaluating GPT3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical. AJNR Am J Neuroradiol. 2024;45(2):244-248. [FREE Full text] [doi: 10.3174/ajnr.A8102] [Medline: 38238092]
40. Zhu Y, Yang X, Wu Y, Zhang W. Leveraging summary guidance on medical report summarization. IEEE J Biomed Health Inform. 2023;27(10):5066-5075. [doi: 10.1109/JBHI.2023.3304376] [Medline: 37566507]
41. Wang H, Wu W, Dou Z, He L, Yang L. Performance and exploration of ChatGPT in medical examination, records and education in Chinese: pave the way for medical AI. Int J Med Inform. 2023;177:105173. [doi: 10.1016/j.ijmedinf.2023.105173] [Medline: 37549499]
42. Yan A, McAuley J, Lu X, Du J, Chang EY, Gentili A, et al. RadBERT: Adapting transformer-based language models to radiology. Radiol Artif Intell. 2022;4(4):e210258. [FREE Full text] [doi: 10.1148/ryai.210258] [Medline: 35923376]
43. Zhao S, Li Q, Yang Y, Wen J, Luo W. From softmax to nucleusmax: A novel sparse language model for Chinese radiology report summarization. ACM Trans Asian Low Resour Lang Inf Process. 2023;22(6):1-21. [doi: 10.1145/3596219]
44. Cai P, Liu F, Bajracharya A, Sills J, Kapoor A, Liu W, et al. Generation of patient after-visit summaries to support physicians. 2022. Presented at: Proceedings of the 29th International Conference on Computational Linguistics; October 1, 2022:6234-6247; Gyeongju, Republic of Korea. URL: https://aclanthology.org/2022.coling-1.544
45. Hartman V, Campion TR. A day-to-day approach for automating the hospital course section of the discharge summary. AMIA Jt Summits Transl Sci Proc. 2022;2022:216-225. [FREE Full text] [Medline: 35854728]
46. Dai S, Wang Q, Lyu Y, Zhu Y. BDKG at MEDIQA 2021: System report for the radiology report summarization task. 2021. Presented at: Proceedings of the 20th Workshop on Biomedical Language Processing; June 1, 2021:103-111; Virtual meeting.
47. Van Veen D, Van Uden C, Attias M, Pareek A, Bluethgen C, Polacin M, et al. RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. 2023. Presented at: The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks; July 1, 2023:449-460; Toronto, ON.
48. Chuang YN, Tang R, Jiang X, Hu X. SPeC: A soft prompt–based calibration on performance variability of large language model in clinical notes summarization. J Biomed Inform. 2024;151:104606. [FREE Full text] [doi: 10.1016/j.jbi.2024.104606] [Medline: 38325698]
49. Gao Y, Miller T, Xu D, Dligach D, Churpek MM, Afshar M. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. Proc Int Conf Comput Ling. 2022;2022:2979-2991. [FREE Full text] [Medline: 36268128]

50. Mehta P, Majumder P. Effective aggregation of various summarization techniques. Inf Process Manag. 2018;54(2):145-158. [doi: 10.1016/j.ipm.2017.11.002]
51. Yu H, Liu J, Zhang X, Wu J, Cui P. A survey on evaluation of out-of-distribution generalization. ArXiv. Preprint posted on March 4, 2024. 2024. [FREE Full text] [doi: 10.48550/arXiv.2403.01874]
52. Goldstein A, Shahar Y. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. J Biomed Inform. 2016;61:159-175. [doi: 10.1016/j.jbi.2016.03.022] [Medline: 27039119]
53. Fabbri A, Kry?ci?ski W, McCann B, Xiong C, Socher R, Radev D. SummEval: Re-evaluating summarization evaluation. Trans Assoc Comput Linguist. 2021;9(2):391-409. [FREE Full text] [doi: 10.1162/tacl_a_00373]
54. Lenharo M. The testing of AI in medicine is a mess. Here's how it should be done. Nature. 2024;632(8026):722-724. [doi: 10.1038/d41586-024-02675-0] [Medline: 39169244]
55. Yang J, Soltan AAS, Clifton DA. Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening. NPJ Digit Med. 2022;5(1):69. [doi: 10.1038/s41746-022-00614-9] [Medline: 35672368]
56. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. [FREE Full text] [doi: 10.1186/s12916-019-1426-2] [Medline: 31665002]
57. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
58. Roberts K. Large language models for reducing clinicians' documentation burden. Nat Med. 2024;30(4):942-943. [doi: 10.1038/s41591-024-02888-w] [Medline: 38561439]
59. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. Patterns. 2021;2(10):100347. [FREE Full text] [doi: 10.1016/j.patter.2021.100347] [Medline: 34693373]
60. Röösli E, Bozkurt S, Hernandez-Boussard T. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. Sci Data. 2022;9(1):24. [FREE Full text] [doi: 10.1038/s41597-021-01110-7] [Medline: 35075160]
61. Gershanik EF, Lacson R, Khorasani R. Critical finding capture in the impression section of radiology reports. AMIA Annu Symp Proc. 2011;2011:465-469. [FREE Full text] [Medline: 22195100]
62. Supriyono, Wibawa AP, Suyono, Kurniawan F. A survey of text summarization: techniques, evaluation and challenges. J. Nat. Lang. Process. 2024;7:100070. [doi: 10.1016/j.nlp.2024.100070]
63. Shakil H, Farooq A, Kalita J. Abstractive text summarization: State of the art, challenges, and improvements. Neurocomputing. 2024;603:128255. [doi: 10.1016/j.neucom.2024.128255]
64. Ryu AJ, Magnuson DR, Kingsley TC. Why mayo clinic is embracing the cloud and what this means for clinicians and researchers. Mayo Clin Proc Innov Qual Outcomes. 2021;5(6):969-973. [FREE Full text] [doi: 10.1016/j.mayocpiqo.2021.08.010] [Medline: 34632298]
65. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C. Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. J Med Internet Res. 2019;21(5):e13484. [FREE Full text] [doi: 10.2196/13484] [Medline: 31152528]
66. 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):60-69. [doi: 10.1038/s41591-024-03425-5] [Medline: 39779929]
67. van der Lee C, Gatt A, van Miltenburg E, Krahmer E. Human evaluation of automatically generated text: current trends and best practice guidelines. Comput Speech Lang. 2021;67:101151. [doi: 10.1016/j.csl.2020.101151]
68. Krishna K, Bransom E, Kuehl B, Iyyer M, Dasigi P, Cohan A, et al. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. 2023. Presented at: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics; May 1, 2023:1650-1669; Dubrovnik, Croatia.
69. Liu Y, Fabbri A, Liu P, Zhao Y, Nan L, Han R, et al. Revisiting the gold standard: grounding summarization evaluation with robust human evaluation. 2023. Presented at: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); July 1, 2023:4140-4170; Toronto, Canada.
70. Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto T. Benchmarking large language models for news summarization. Trans Assoc Comput Linguist. 2024;12:39-57. [FREE Full text] [doi: 10.1162/tacl_a_00632]
71. 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. [FREE Full text] [doi: 10.1038/s41746-024-01258-7] [Medline: 39333376]
72. Song Y, Wang G, Li S, Lin BY. The good, the bad, and the greedy: Evaluation of LLMs should not ignore non-determinism. ArXiv. Preprint posted on July 15, 2024. 2024. [FREE Full text] [doi: 10.48550/arXiv.2407.10457]
73. Cabitza F, Campagner A, Soares F, García de Guadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. 2021;208:106288. [FREE Full text] [doi: 10.1016/j.cmpb.2021.106288] [Medline: 34352688]
74. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health. 2020;2(9):e489-e492. [FREE Full text] [doi: 10.1016/S2589-7500(20)30186-2] [Medline: 32864600]

75. Reddy S. Evaluating large language models for use in healthcare: A framework for translational value assessment. Informatics in Medicine Unlocked. 2023;41:101304. [doi: 10.1016/j.imu.2023.101304]
76. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. ArXiv. Preprint posted on July 12, 2022. 2021. [FREE Full text]
77. de Hond A, Leeuwenberg T, Bartels R, van Buchem M, Kant I, Moons KG, et al. From text to treatment: The crucial role of validation for generative large language models in health care. Lancet Digit Health. 2024;6(7):e441-e443. [FREE Full text] [doi: 10.1016/S2589-7500(24)00111-0] [Medline: 38906607]
78. Article 9: Risk management system. EU Artificial Intelligence Act. URL: https://artificialintelligenceact.eu/article/9/ [accessed 2025-02-14]
79. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans Inf Syst. 2025;43(2):1-55. [doi: 10.1145/3703155]
80. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1-38. [doi: 10.1145/3571730]
81. Asgari E, Montaña-Brown N, Dubois M, Khalil S, Balloch J, Pimenta D. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. medRxiv. 2024. [doi: 10.1101/2024.09.12.24313556]
82. Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, Dieng A, et al. A toolbox for surfacing health equity harms and biases in large language models. Nat Med. 2024;30(12):3590-3600. [doi: 10.1038/s41591-024-03258-2] [Medline: 39313595]
83. Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, et al. Bias and fairness in large language models: a survey. Comput Linguist. 2024;50:1097-1179. [FREE Full text] [doi: 10.1162/coli_a_00524]
84. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024;6(1):e12-e22. [FREE Full text] [doi: 10.1016/S2589-7500(23)00225-X] [Medline: 38123252]
85. PhysioNet. URL: https://physionet.org/news/post/gpt-responsible-use [accessed 2025-04-16]
86. Ong JCL, Chang SY, William W, Butte AJ, Shah NH, Chew LST, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health. 2024;6(6):e428-e432. [FREE Full text] [doi: 10.1016/S2589-7500(24)00061-X] [Medline: 38658283]
87. Miller K. De-identifying medical patient data doesn't protect our privacy. Stanford University Human-Centered Artificial Intelligence. URL: https://hai.stanford.edu/news/de-identifying-medical-patient-data-doesnt-protect-our-privacy?utm_source=chatgpt.com [accessed 2025-04-16]
88. Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: techniques and applications. Comput Biol Med. 2023;158:106848. [FREE Full text] [doi: 10.1016/j.compbiomed.2023.106848] [Medline: 37044052]
89. Pollock D, Peters MDJ, Khalil H, McInerney P, Alexander L, Tricco AC, et al. Recommendations for the extraction, analysis, and presentation of results in scoping reviews. JBI Evid Synth. 2023;21(3):520-532. [FREE Full text] [doi: 10.11124/JBIES-22-00123] [Medline: 36081365]

Abbreviations

API: application programming interface
EHR: electronic health record
IRB: institutional review board
LLM: large language model
MIMIC: Medical Information Mart for Intensive Care
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

Edited by T de Azevedo Cardoso; submitted 20.11.24; peer-reviewed by D Chrimes, R Lin, S Sivarajkumar, P Yu, C Lee; comments to author 27.01.25; revised version received 21.02.25; accepted 12.03.25; published 15.05.25

Please cite as:
Bednarczyk L, Reichenpfader D, Gaudet-Blavignac C, Ette AK, Zaghir J, Zheng Y, Bensahla A, Bjelogrlic M, Lovis C
Scientific Evidence for Clinical Text Summarization Using Large Language Models: Scoping Review
J Med Internet Res 2025;27:e68998
URL: https://www.jmir.org/2025/1/e68998
doi: 10.2196/68998
PMID: