The Evolution of Clinical Intelligence Through GenAI Co-pilots: A Systematic Review and Thematic Synthesis

Li-Chen Cheng[1], Chia-Yu Hung[2], Te-Nien Chien[3]*
Department of Information and Finance Management, National Taipei University of Technology, Taipei, Taiwan[1]
College of Management, National Taipei University of Technology, Taipei, Taiwan[2, 3]

***Abstract*—Generative artificial intelligence (GenAI), large language models (LLMs), and multimodal foundation models are rapidly transforming healthcare by extending artificial intelligence beyond traditional predictive analytics toward collaborative clinical intelligence. Recent advances have enabled applications in clinical decision support, medical documentation, workflow optimization, patient communication, and personalized care planning. However, the existing literature remains fragmented across technical evaluations, specialty-specific applications, and governance discussions, resulting in a limited understanding of how these technologies collectively function within clinical environments. This study conducted a systematic review and thematic synthesis to examine the emerging role of GenAI as a clinical co-pilot in healthcare. Following the PRISMA 2020 framework, literature was retrieved from Web of Science, PubMed, and Scopus databases. A total of 3,124 records were identified, and 41 studies published between 2023 and March 2026 met the eligibility criteria and were included in the final qualitative synthesis. Quality assessment indicated that 87.8% of the included studies were classified as moderate or high quality, providing a robust methodological basis for the thematic synthesis. Thematic analysis revealed three interrelated domains underlying the evolution of GenAI-enabled clinical intelligence: 1) Perception and Fact Anchoring, involving multimodal data integration, retrieval-augmented generation (RAG), and domain-specific medical intelligence; 2) Clinical Agency and Collaboration, encompassing ambient digital scribing, agentic clinical decision support, workflow augmentation, and personalized care; and 3) Governance and Responsibility, including human-in-the-loop oversight, privacy protection, regulatory compliance, fairness, and trustworthy AI practices. Based on these findings, a conceptual Clinical Co-pilot Framework is proposed to position GenAI as a collaborative partner that supports clinicians rather than replaces them. The framework provides a conceptual basis for future empirical validation and may help inform the responsible implementation of GenAI in healthcare.**

Keywords*—Generative artificial intelligence; clinical co-pilot; large language models; clinical decision support; human-AI collaboration

I. Introduction

A. Background

The rapid advancement of generative artificial intelligence (GenAI) has significantly transformed digital healthcare and clinical decision support [1]. Since the emergence of advanced large language models (LLMs) and multimodal foundation models, including GPT-4, Med-PaLM, Claude, and domain-specific healthcare models, healthcare organizations have increasingly explored the integration of GenAI into clinical environments [2-6]. Compared with traditional AI systems focused on prediction and classification, contemporary GenAI technologies offer broader capabilities in contextual reasoning, language generation, multimodal interpretation, and human-AI interaction [7].

Recent developments have enabled new forms of clinical support, including electronic health record (EHR) summarization, evidence retrieval, differential diagnosis generation, clinical documentation, and physician communication support [8-14]. The rapid growth of healthcare data and clinical complexity has further accelerated interest in these technologies [15]. Consequently, healthcare institutions are increasingly investigating GenAI as a collaborative tool to augment clinical reasoning, improve workflow efficiency, and reduce administrative burden [16-20].

Recent literature suggests that healthcare applications of GenAI are transitioning from proof-of-concept studies toward real-world clinical implementation [21-23]. Current applications include clinical documentation, retrieval-augmented decision support, AI-assisted triage, personalized treatment planning, multimodal evidence synthesis, and patient education [24-28]. At the same time, concerns regarding hallucination, bias, transparency, accountability, privacy, and regulatory compliance have emerged as major challenges for responsible deployment [27, 28].

B. Research Motivation

Traditional healthcare artificial intelligence systems have historically emphasized automation-oriented objectives, including disease prediction, risk stratification, image classification, and algorithmic screening [16, 29, 30]. Although these technologies demonstrated promising benchmark performance, many failed to achieve sustainable implementation in real-world clinical environments due to poor workflow integration, limited interpretability, clinician distrust, insufficient adaptability, and limited support for complex collaborative decision-making.

Unlike earlier predictive AI systems, GenAI introduces a different interaction paradigm between clinicians and intelligent systems. Rather than functioning solely as automated prediction

*Corresponding author.

engines, GenAI technologies increasingly operate as collaborative assistants capable of supporting clinical reasoning, information synthesis, workflow coordination, and communication under physician supervision [9, 31]. This shift reflects a broader movement from automation-centered artificial intelligence toward collaboration-centered clinical intelligence.

Within this emerging paradigm, the concept of the "clinical co-pilot" has gained increasing relevance. Inspired by the co-pilot model in aviation, this framework conceptualizes GenAI as an augmentative partner designed to support, rather than replace, healthcare professionals [16, 32]. Physicians remain the final decision-makers and accountable authorities, while GenAI assists by enhancing situational awareness, reducing cognitive burden, and improving access to contextual medical knowledge [16, 33, 34].

This framework is particularly important because healthcare decision-making involves uncertainty, ethical responsibility, contextual judgment, and interpersonal communication [33, 35]. Fully autonomous clinical decision-making remains difficult to implement safely in high-stakes environments. Therefore, collaborative human-AI systems may represent a more feasible and responsible pathway toward intelligent healthcare transformation [10, 34].

Current literature on GenAI in healthcare remains fragmented across technical applications, workflow integration, ethical governance, implementation barriers, and clinical evaluation. Many reviews focus narrowly on ChatGPT performance, benchmark testing, or isolated healthcare applications without systematically conceptualizing how GenAI may function as a collaborative partner within healthcare ecosystems [15, 32, 36]. Therefore, a comprehensive synthesis integrating technological capabilities, clinical workflows, governance mechanisms, and human-AI collaboration remains necessary [16].

C. Problem Statement and Research Gap

The rapid advancement of GenAI, LLMs, and multimodal foundation models has accelerated the adoption of AI technologies across diverse healthcare settings [7]. Recent studies have demonstrated promising applications in clinical decision support, medical documentation, patient communication, diagnostic assistance, treatment planning, and workflow optimization [37-39].

Despite this progress, the literature remains fragmented across disciplines and application domains. Many studies focus on evaluating specific models such as ChatGPT, GPT-4, and Med-PaLM using benchmark datasets such as MedQA and MMLU [15, 35, 40, 41], while others examine applications within specific specialties including radiology, oncology, pathology, emergency medicine, and nursing practice [10,29]. At the same time, substantial attention has been devoted to challenges related to hallucination, explainability, privacy, bias, transparency, and regulatory oversight [9, 20, 29, 42-44].

Although these studies provide valuable insights, relatively few have examined how technological capabilities, clinical workflows, and governance mechanisms collectively contribute to human-AI collaboration in healthcare. Existing reviews remain largely focused on model performance or isolated applications and rarely integrate these perspectives within a unified conceptual framework [18, 45].

As healthcare AI continues to evolve from task-specific automation toward collaborative intelligence, there is a growing need to understand how GenAI supports clinical workflows while operating within ethical, legal, and organizational boundaries [46, 47]. Therefore, this study systematically reviews current evidence and develops a clinical co-pilot framework that integrates technology, workflow, and governance perspectives to explain the evolution of collaborative human-AI healthcare systems.

D. Research Objectives

This study aims to systematically review and synthesize current literature regarding the applications of GenAI in healthcare through the conceptual perspective of the clinical co-pilot framework.

Specifically, this study seeks to achieve the following objectives:

- To systematically identify and classify major application domains of GenAI in clinical decision support and healthcare collaboration.
- To examine how GenAI technologies support clinical reasoning, multimodal information integration, workflow coordination, and healthcare communication.
- To analyze the implementation challenges and governance concerns associated with GenAI deployment, including hallucination risk, accountability, bias, privacy protection, and regulatory compliance.
- To propose a conceptual framework describing GenAI as a collaborative clinical co-pilot operating under human supervision.
- To explore how healthcare AI is evolving from automation-centered systems toward collaboration-centered augmented clinical intelligence.
- To identify current research gaps and future directions for responsible human-AI collaboration in healthcare environments.

The findings suggest that the clinical co-pilot model represents a feasible and responsible pathway for integrating GenAI into healthcare systems while preserving human oversight, ethical governance, and clinician-centered decision-making.

II. METHODS

A. Study Design

This study was designed as a systematic review with thematic synthesis to examine how GenAI has been applied as a collaborative clinical co-pilot in healthcare and clinical decision support. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 (PRISMA 2020) framework to enhance transparency, reproducibility, and methodological rigor [48, 49].

The systematic review component was used to identify, screen, and select relevant studies through a structured and reproducible search process. The thematic synthesis component was used to integrate findings across heterogeneous studies and to develop a conceptual framework explaining the role of GenAI as a clinical co-pilot. This combined approach was appropriate because the included literature covered diverse study designs, clinical settings, model architectures, and implementation contexts.

The review focused on studies published from January 2023 to March 2026, corresponding to the period in which advanced LLMs and multimodal GenAI systems rapidly entered healthcare research and clinical implementation discussions. The unit of analysis was the individual scholarly article.

B. Search Strategy

A structured literature search strategy was developed to identify studies examining the application of GenAI technologies in healthcare settings, with particular emphasis on clinical decision support, healthcare workflows, and human-AI collaboration. The search strategy was designed to ensure comprehensive coverage of the rapidly expanding literature on LLMs, multimodal AI systems, and their integration into clinical practice.

The literature search was conducted using three major academic databases: Web of Science Core Collection, PubMed, and Scopus. These databases were selected because they provide extensive coverage of peer-reviewed research in medicine, healthcare informatics, artificial intelligence, computer science, and interdisciplinary healthcare technologies. The search covered publications from January 2023 to March 2026, corresponding to the period following the emergence and widespread adoption of advanced LLMs and multimodal GenAI systems.

This specific time frame was intentionally selected because the public release of ChatGPT in late 2022 marked a major turning point in the development of GenAI technologies. Following this milestone, healthcare research experienced a rapid increase in studies involving LLMs, multimodal foundation models, retrieval-augmented generation (RAG), and agentic AI applications. Consequently, the period from 2023 to March 2026 captures the early evolution of GenAI from experimental demonstrations toward clinical implementation, workflow integration, and healthcare governance. This timeframe therefore provides an appropriate foundation for examining the emergence of GenAI as a clinical co-pilot in healthcare environments.

To maximize retrieval sensitivity while maintaining conceptual relevance, the search strategy incorporated three core domains:

- GenAI technologies, including LLMs, foundation models, and multimodal AI systems;
- Clinical intelligence and healthcare support functions, including clinical decision support, clinical reasoning, workflow assistance, and human-AI collaboration;
- Healthcare environments, including hospitals, clinical practice, healthcare delivery systems, and patient care settings.

Boolean operators (AND, OR), truncation symbols (*), and phrase searching were employed to maximize retrieval sensitivity while minimizing irrelevant records. Database-specific search syntax was adapted for Web of Science, PubMed, and Scopus while maintaining conceptual consistency across all information sources.

The database search yielded a total of 3,124 records, including 957 records from Web of Science, 1,048 records from PubMed, and 1,119 records from Scopus.

The core search syntax was defined as follows:

("generative artificial intelligence"

OR "generative AI"OR "large language model*"OR LLM*OR ChatGPT\OR "foundation model*"OR "multimodal large language model*")

AND

("clinical decision support"OR "clinical decision support system*"OR CDSS OR "clinical decision-making"OR "clinical workflow" OR "clinical co-pilot" OR "medical co-pilot" OR "human-AI collaboration" OR "clinical reasoning" OR "AI agent*"OR "retrieval-augmented generation")

AND

(healthcare OR medicine OR clinical OR hospital OR "patient care")

For the Web of Science Core Collection, searches were conducted using the Topic (TS) field, which includes titles, abstracts, author keywords, and Keywords Plus. In Scopus, searches were performed using the TITLE-ABS-KEY field, while equivalent search syntax was adapted for PubMed. Minor database-specific adjustments were applied where necessary while maintaining conceptual consistency across all information sources.

The search strategy was intentionally designed to capture studies examining the application of GenAI technologies within clinical environments, particularly those involving clinical decision support, healthcare workflows, human-AI collaboration, intelligent clinical assistance, and responsible AI implementation. The broad search approach ensured comprehensive coverage of the rapidly evolving literature while allowing subsequent screening and thematic synthesis to identify the most relevant evidence related to emerging forms of AI-assisted clinical intelligence.

Importantly, the search strategy was intentionally constructed using broad concepts related to GenAI, clinical decision support, healthcare workflows, and human-AI collaboration rather than predefined theoretical categories. The concept of the clinical co-pilot was not used as a mandatory inclusion criterion and did not predetermine the thematic structure of the review.

Instead, the clinical co-pilot framework emerged inductively through the thematic synthesis process as recurring patterns of technological capabilities, workflow integration, and governance mechanisms were identified across the included studies.

The database search process and retrieved records are summarized in Table I.

The initial database search yielded 3,124 records, including 957 records from Web of Science, 1,048 records from PubMed, and 1,119 records from Scopus. Following record export and consolidation, duplicate and non-eligible records were removed prior to screening. The subsequent study selection process was conducted according to the PRISMA 2020 guidelines and involved title and abstract screening, full-text eligibility assessment, and thematic synthesis. The detailed study selection process is presented in Fig. 1.

TABLE I. LITERATURE SEARCH SOURCES AND RETRIEVAL RESULTS

Database	Search Field	Records Retrieved
Web of Science	Topic Search	957
PubMed	Title / Abstract Search	1,048
Scopus	TITLE-ABS-KEY Search	1,119
Total		3,124

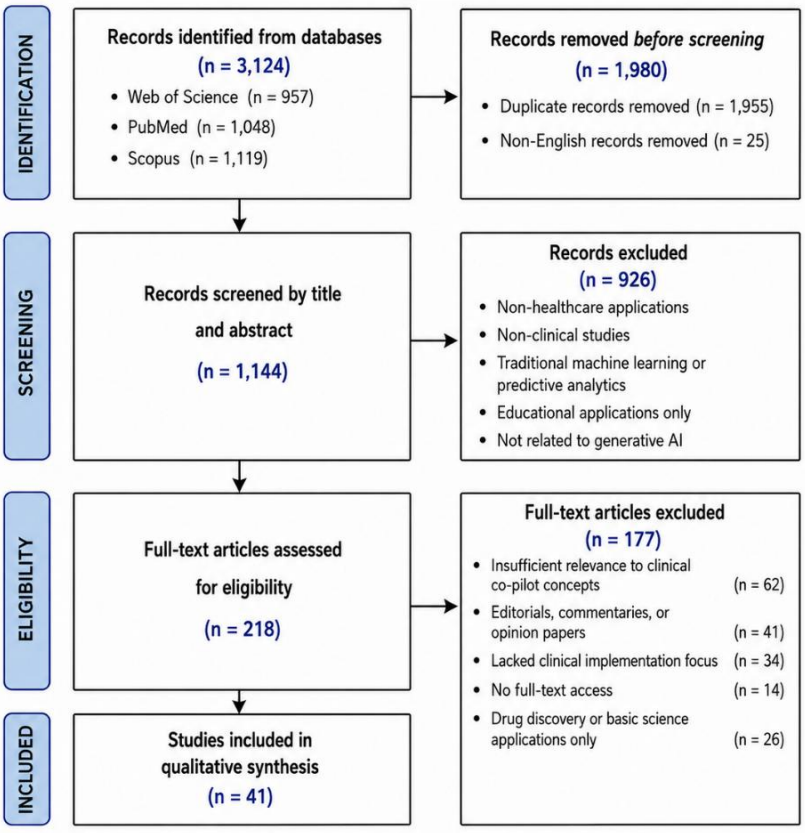

Fig. 1. PRISMA 2020 flow diagram.

C. Inclusion and Exclusion Criteria

Eligibility criteria were defined before screening to ensure consistency in study selection. Studies were included if they met all of the following criteria:

- The study focused on GenAI, LLMs, multimodal generative models, RAG, or related foundation model technologies.
- The study was situated in healthcare, medicine, hospital practice, clinical care, or patient-related decision-making contexts.
- The study discussed or evaluated applications relevant to clinical decision support, clinical workflow assistance, physician decision-making, patient data interpretation, treatment planning, documentation support, or human-AI collaboration.
- The article provided sufficient methodological, conceptual, or application-related information for thematic extraction.
- The article was published in English between January 2023 and March 2026.

Studies were excluded if they met any of the following criteria:

- The study focused only on traditional machine learning or predictive modeling without a GenAI component.
- The study was unrelated to healthcare or clinical practice.

- The study focused solely on medical education, student learning, or examination performance without clear clinical decision support relevance.
- The article was an editorial, commentary, news item, short opinion piece, or letter without sufficient analytical content.
- The study focused exclusively on drug discovery, molecular generation, or basic biomedical research without connection to clinical decision support or healthcare delivery.
- Full text was unavailable, or the article lacked sufficient information for data extraction.

These criteria were designed to keep the review focused on GenAI as a collaborative clinical tool rather than on general AI applications in healthcare.

D. Study Selection

The study selection process followed the PRISMA 2020 flow structure, including identification, screening, eligibility assessment, and final inclusion. First, all records retrieved from the selected databases were compiled. Duplicate records were identified and removed before screening. Next, titles and abstracts were reviewed to exclude clearly irrelevant studies. Articles that appeared potentially relevant were then retrieved for full-text assessment.

During full-text review, each article was evaluated against the predefined inclusion and exclusion criteria. Studies were excluded at this stage if they did not address GenAI, lacked clinical decision support relevance, focused only on education or non-clinical settings, or provided insufficient analytical content. Reasons for exclusion were documented to support PRISMA transparency.

A total of 3,124 records were identified through database searching. Following duplicate removal, title and abstract screening, and full-text eligibility assessment, 41 studies met the inclusion criteria and were included in the final qualitative synthesis. Details of the study selection process are presented in Fig. 1.

E. Data Extraction and Thematic Synthesis

A structured data extraction form was developed to collect key information from each included study. Extracted items included author, publication year, country or region, clinical domain, study design, data source, GenAI model type, clinical application, evaluation method, reported benefits, limitations, governance concerns, and relevance to the clinical co-pilot framework.

The extracted data were organized into two main evidence tables. The first table summarized the characteristics of included studies, including publication year, clinical domain, model type, data source, and study design. The second table classified studies according to their clinical co-pilot functions, including perception and fact anchoring, clinical agency and collaboration, and governance and responsibility boundaries.

Thematic synthesis was conducted using an iterative coding process. First, open coding was used to identify recurring concepts across studies, such as EHR summarization, RAG, ambient documentation, hallucination, accountability, and human-in-the-loop oversight. Second, axial coding was used to group related codes into broader thematic categories. Third, selective coding was used to develop an integrated conceptual framework explaining GenAI as a clinical co-pilot.

Three major themes were identified:

- Perception and fact anchoring: GenAI capabilities for understanding, integrating, retrieving, and contextualizing clinical information.
- Clinical agency and collaboration: GenAI applications that support documentation, diagnosis, treatment planning, workflow coordination, and clinical communication.
- Governance and responsibility boundaries: institutional, ethical, legal, and human oversight mechanisms required for safe clinical deployment.

This thematic synthesis approach allowed the review to move beyond a descriptive summary of applications and toward a conceptual interpretation of how GenAI may function as a responsible and collaborative partner in healthcare.

F. PRISMA Reporting and Methodological Rigor

To strengthen methodological transparency, the findings followed the PRISMA 2020 checklist throughout the search, screening, eligibility assessment, and reporting process. The review explicitly documented the information sources, search strategy, eligibility criteria, selection process, data extraction procedure, and synthesis method.

The PRISMA flow diagram was used to visually summarize the study selection process. Full-text exclusion reasons were categorized and reported to clarify how the final evidence base was constructed. The inclusion and exclusion criteria were applied consistently during screening to reduce selection bias. Because the included studies were expected to be heterogeneous in design, ranging from empirical evaluations and case studies to conceptual frameworks and systematic reviews, a meta-analysis was not conducted. Instead, thematic synthesis was selected as the most appropriate method for integrating diverse evidence types. This approach enabled the identification of cross-cutting themes related to clinical usefulness, implementation feasibility, governance requirements, and human-AI collaboration.

To further enhance rigor, the review distinguished between technical capability, clinical workflow relevance, and governance implications. This distinction helped prevent overgeneralization from benchmark-based performance studies to real-world clinical practice. The final synthesis therefore emphasizes not only what GenAI can do in healthcare, but also how it can be responsibly integrated into clinician-centered decision-making environments.

The review was not prospectively registered in PROSPERO or other systematic review registries because of the rapidly evolving nature of the GenAI evidence base and the conceptual objective of thematic synthesis. The study focused on

conceptual integration and framework development rather than quantitative meta-analysis or intervention effect estimation.

G. Quality Assessment

A methodological quality assessment was conducted to evaluate the rigor and credibility of the included studies.

Given the heterogeneous nature of the evidence base, including empirical studies, implementation reports, conceptual frameworks, and review articles, a simplified quality appraisal framework was adopted. The quality appraisal criteria were adapted from the Mixed Methods Appraisal Tool (MMAT) and previous healthcare artificial intelligence review studies to accommodate the diverse methodological characteristics of the included literature.

Studies were evaluated according to five criteria:

- Clarity of research objectives
- Methodological transparency
- Description of AI technologies
- Clinical relevance
- Governance considerations

Each criterion was evaluated using a three-point scale (0 = not addressed, 1 = partially addressed, and 2 = clearly addressed), resulting in a maximum score of 10 points for each study. Studies scoring 8–10 points were classified as high quality, 5–7 points as moderate quality, and 0–4 points as low quality. Two reviewers independently performed the quality assessment, and disagreements were resolved through discussion and consensus. The quality assessment results of the included studies are summarized in Table II.

TABLE II. SUMMARY OF QUALITY ASSESSMENT RESULTS

Quality Level	Score Range	Number (n)	Percentage (%)
High	8-10	15	36.6%
Moderate	5-7	21	51.2%
Low	0-4	5	12.2%
Total		41	100%

Note: Five quality criteria were evaluated. Total scores ranged from 0 to 10, with higher scores indicating greater methodological rigor.

Overall, most studies were classified as moderate-to-high quality (87.8%), suggesting that the current evidence base provides a reasonably robust foundation for thematic synthesis and supports the development of the proposed conceptual framework. The quality assessment was used to evaluate study rigor and to support the interpretation of thematic findings rather than to exclude studies from the review. No study was excluded solely on the basis of quality assessment results. Given the exploratory and conceptual nature of this study, all eligible studies were retained to ensure comprehensive coverage of emerging evidence related to GenAI clinical co-pilots.

III. RESULTS

A. Overview of Included Studies

A total of 41 studies were included in the final qualitative synthesis following the PRISMA 2020 screening process. These studies were selected from an initial pool of 3,124 records retrieved from Web of Science, PubMed, and Scopus. The included studies were published between 2023 and 2026, reflecting the rapid expansion of GenAI research following the emergence of advanced foundation models and LLMs. The overall publication trend demonstrated a substantial increase in healthcare-related GenAI studies after 2023, particularly in clinical decision support, workflow automation, and healthcare governance applications.

To provide an overview of the included literature, the characteristics of the 41 selected studies were summarized according to publication year, study type, AI technology, clinical focus, and thematic domain (Table III). This descriptive analysis provides contextual insights into the evolving landscape of GenAI research in healthcare and serves as a foundation for the subsequent thematic synthesis. The characteristics of the 41 included studies, including study type, main technology, and primary application domains, are summarized in Table V (Appendix).

TABLE III. OVERVIEW OF INCLUDED STUDIES

Category	Subcategory	Frequency (n)	Percentage (%)
Publication Year	2023	6	14.6
	2024	12	29.3
	2025	18	43.9
	2026	5	12.2
Study Type	Empirical Evaluation	15	36.6
	Framework / Conceptual Study	8	19.5
	Case Study / Implementation	10	24.4
	Review / Perspective	8	19.5
AI Technology	Large Language Models (LLMs)	17	41.5
	Retrieval-Augmented Generation (RAG)	7	17.1
	Multimodal LLMs (MLLMs)	9	22.0
	Agentic AI / AI Agents	4	9.8
	Other GenAI Approaches	4	9.8
Clinical Focus	Clinical Decision Support	14	34.1
	Clinical Documentation	7	17.1
	Personalized Care Planning	5	12.2
	Workflow Optimization	6	14.6
	Governance and Ethics	9	22.0
Thematic Domain	Perception and Fact Anchoring	13	31.7
	Clinical Agency and Collaboration	15	36.6
	Governance and Responsibility	13	31.7

Note: Publications published between January 2023 and March 2026 were included in the review.

The studies were categorized according to their primary research focus, dominant AI technology, and thematic contribution. Each study was assigned to a single primary category within each dimension to facilitate descriptive analysis and thematic synthesis.

The descriptive statistics indicate a rapid increase in healthcare-related GenAI research following the emergence of advanced LLMs and multimodal foundation models. To further illustrate the temporal distribution of the included studies, the annual publication trend is presented in Fig. 2.

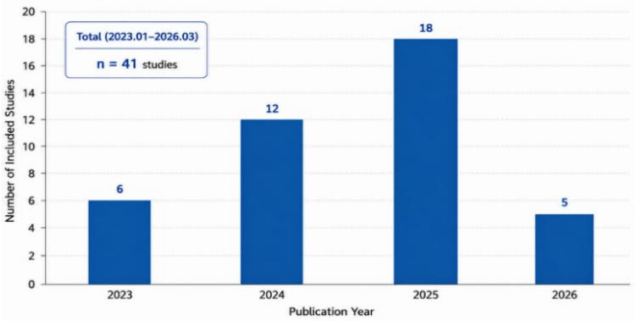

Fig. 2. Publication trends of included studies (2023.01–2026.03).

The reviewed studies originated from diverse disciplinary backgrounds, including healthcare informatics, clinical medicine, digital health, computer science, and interdisciplinary artificial intelligence research. Most studies focused on hospital-based or clinical healthcare environments, while a smaller number examined broader healthcare management systems, patient communication, or institutional governance frameworks.

In terms of technological focus, GPT-based architectures and LLMs were the most frequently discussed GenAI technologies. Commonly referenced systems included GPT-4, ChatGPT, Med-PaLM, Claude, and multimodal large language models (MLLMs). Several studies also explored RAG, domain-specific medical foundation models, and multimodal clinical reasoning systems integrating electronic health records, medical imaging, and laboratory data [3, 5, 50, 51].

The included studies covered multiple clinical application domains. The most frequently identified applications included clinical documentation and summarization, clinical decision support, treatment planning, ambient scribing, patient data interpretation, evidence retrieval, and workflow coordination. Several studies also examined physician acceptance, trust calibration, human-AI collaboration, and governance mechanisms associated with clinical AI deployment.

Methodologically, the reviewed literature was heterogeneous in design. The evidence base included empirical evaluations, implementation case studies, benchmark comparisons, conceptual frameworks, observational studies, expert commentaries, and systematic reviews. Due to the diversity of study designs and evaluation metrics, a quantitative meta-analysis was not conducted. Instead, thematic synthesis was used to identify recurring conceptual patterns across the literature.

Beyond publication growth and technological advancement, thematic synthesis revealed three major domains underlying the evolution of GenAI clinical co-pilots: 1) Perception and Fact Anchoring, 2) Clinical Agency and Collaboration, and 3) Governance and Responsibility. Together, these domains suggest a shift from automation-oriented healthcare AI toward collaborative clinical intelligence systems.

B. Conceptual Framework of GenAI as a Clinical Co-Pilot

The thematic synthesis identified a conceptual shift from automation-oriented healthcare AI toward collaborative clinical intelligence. To explain this transition, a conceptual framework of GenAI as a clinical co-pilot was developed based on the three major themes emerging from the reviewed literature: Perception and Fact Anchoring, Clinical Agency and Collaboration, and Governance and Responsibility.

These three domains collectively describe how GenAI systems transform heterogeneous healthcare data into clinically meaningful intelligence, support healthcare professionals through collaborative reasoning and workflow augmentation, and operate within ethical, legal, and organizational governance structures. Rather than functioning as autonomous decision-makers, GenAI systems are increasingly positioned as collaborative partners that augment clinician capabilities while preserving human oversight and accountability.

As illustrated in Fig. 3, the framework begins with diverse healthcare inputs, including electronic health records, medical imaging, laboratory findings, clinical guidelines, and patient histories. Through multimodal reasoning, evidence grounding, and domain-specific intelligence, these inputs are converted into actionable clinical insights that support diagnosis, documentation, communication, care planning, and clinical decision support. The framework is further reinforced by foundational deployment enablers, including interoperable healthcare data, model transparency, continuous evaluation, and regulatory alignment, which collectively support responsible and trustworthy implementation in healthcare environments.

The framework was developed inductively through thematic synthesis and does not represent a predefined theoretical model. The framework begins with diverse clinical inputs originating from electronic health records, medical imaging, laboratory findings, wearable devices, clinical guidelines, and patient histories. GenAI systems synthesize these heterogeneous data sources through multimodal reasoning, RAG, and domain-specific medical intelligence architectures [23, 36]. The resulting outputs support diagnostic reasoning, clinical documentation, workflow optimization, personalized care planning, interdisciplinary communication, and clinician decision support within real-world healthcare environments [16, 19].

The framework conceptualizes GenAI as a collaborative co-pilot rather than an autonomous medical authority. Across the reviewed studies, clinicians consistently remained the final reviewers and accountable decision-makers within AI-assisted healthcare systems [12, 28]. Consequently, governance mechanisms such as human-in-the-loop oversight, transparency, fairness auditing, privacy protection, and regulatory compliance emerged as foundational requirements for responsible deployment [16, 32].

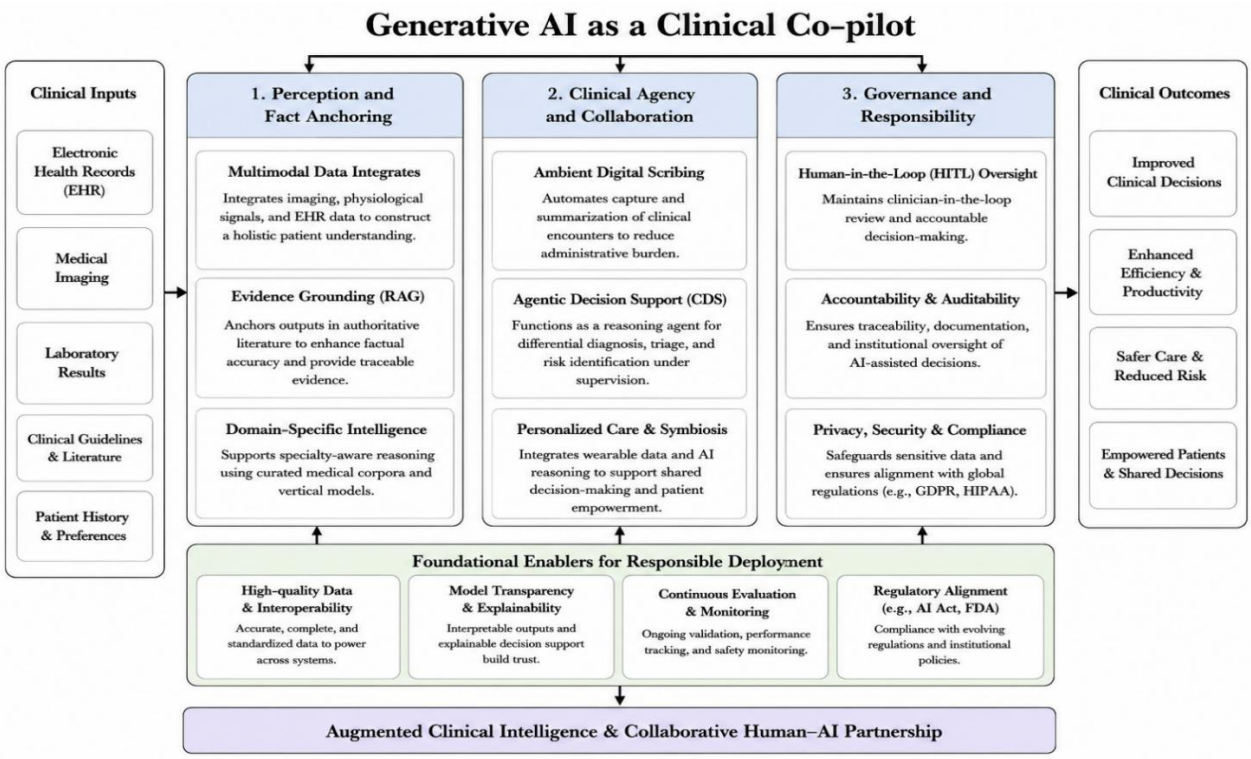

Fig. 3. Clinical Co-pilot framework derived from thematic synthesis of 41 included studies.

TABLE IV. OPTIMIZED THEMATIC CLASSIFICATION OF GENERATIVE AI CLINICAL CO-PILOT FUNCTIONS

Main Theme	Subtheme	AI Co-pilot Role & Functionality	Representative Technologies & Applications	Clinical Significance & Evaluation Metrics	Supporting Studies
Perception and Fact Anchoring	Multimodal Data Integration	Integrates imaging, physiological signals, and EHR data to construct a holistic patient understanding.	MLLMs (e.g., Med-Flamingo, LLaVA-Med), vision-language alignment, multimodal EHR fusion.	Enhances clinical situational awareness; Metrics: Diagnostic accuracy and cross-modal consistency.	[11, 14, 18, 27, 29, 33, 36, 44]
	Evidence Grounding (RAG)	Anchors outputs in authoritative literature to eliminate hallucinations and provide traceable citations.	Retrieval-Augmented Generation (RAG), hybrid retrieval pipelines, evidence-grounded CDS.	Establishes evidence-based trust; Metrics: ACUTE (Traceability, Factuality).	[7, 9, 10, 13, 16, 21, 23, 41, 43, 52]
	Domain-Specific Intelligence	Supports specialty-aware reasoning (e.g., oncology, surgery) using curated medical corpora.	Vertical medical models (BioBERT, ClinicalBERT), disease-scenario association matrices.	Improves specialty-specific precision and relevance; Metrics: MedQA board-style benchmarks.	[6, 9, 11, 14, 21-24, 41, 42, 44, 53, 54]
Clinical Agency and Collaboration	Ambient Digital Scribing	Automates capture and summarization of clinical encounters to reduce administrative burden.	Ambient AI (DAX Co-pilot), SCRIBE framework, automated SOAP note generation.	Reduces physician burnout and documentation time; Metrics: Note completeness and editing time.	[6, 8, 11, 12, 16, 19, 30, 31, 33, 44]
	Agentic Decision Support (Agentic CDS)	Functions as a reasoning "agent" capable of differential diagnosis and active triage under supervision.	AI Agents, virtual triage assistants, interactive diagnostic rationales.	Supports complex reasoning and risk identification; Metrics: AI-SCE (Clinical Skills Assessment).	[13, 15, 16, 18, 19, 23, 28, 33, 35-37, 41, 54, 55]
	Personalized Care & Symbiosis	Integrates wearable data and AI reasoning to support shared decision-making and patient empowerment.	Personalized Patient Care Plans (PPCPs), remote monitoring, health empowerment tools.	Enhances care coordination and patient adherence; Metrics: Patient satisfaction and compliance.	[9, 16, 18-20, 24, 26-30, 36, 44]
Governance and Responsibility	Human-in-the-Loop (HITL) Oversight	Maintains clinicians as the final reviewers and accountable decision-makers.	iLEAP lifecycle model, clinician validation interfaces, supervised CDS protocols.	Preserves diagnostic safety and legal accountability; Metrics: Human-AI decision agreement rates.	[10, 12, 16, 18, 19, 23, 24, 28, 33, 37, 44, 52, 54]
	Privacy, Security, & Compliance	Safeguards sensitive data and ensures alignment with global regulations (GDPR/HIPAA).	Data de-identification, secure infrastructure, quantitative assurance frameworks.	Reduces liability and privacy risks; Metrics: Regulatory compliance and data leakage rates.	[7, 8, 10, 14, 16-19, 23, 28, 32, 35, 37, 42-44, 52, 56]
	Ethics, Fairness, & Trustworthy AI	Mitigates algorithmic bias and promotes transparency and equitable healthcare delivery.	Fairness audits, AVI intelligent filters, explainable AI (XAI), bias mitigation.	Supports health equity and trustworthy deployment; Metrics: Bias reduction rate and explainability scores.	[9, 18, 19, 23, 25, 27, 28, 32, 35, 37, 42-44, 52, 53]

To further synthesize the major themes identified across the included studies, Table IV presents the thematic classification of GenAI clinical co-pilot functions. The classification was developed inductively through thematic synthesis of the included studies and serves as an analytical framework for organizing the identified clinical co-pilot functions. Table IV summarizes the conceptual roles, representative technologies, clinical significance, evaluation considerations, and supporting evidence associated with each thematic domain. The findings suggest that contemporary GenAI systems are evolving beyond isolated automation tools toward integrated collaborative intelligence infrastructures capable of supporting multimodal reasoning, clinical workflow augmentation, and responsible healthcare governance.

C. Perception and Fact Anchoring

The first major thematic domain identified in the findings involved perception and fact anchoring, representing the foundational cognitive layer of GenAI clinical co-pilot systems. This domain focuses on the ability of GenAI to interpret, integrate, contextualize, and retrieve heterogeneous healthcare information in order to support clinical reasoning and evidence-based decision-making.

One of the most frequently discussed capabilities involved multimodal data integration. Several studies described the emergence of multimodal LLMs capable of combining electronic health records, medical imaging, laboratory findings, physiological signals, and patient histories within unified reasoning systems [36]. Unlike conventional language-only architectures, these multimodal systems provide a more comprehensive understanding of patient conditions by integrating diverse clinical modalities into a shared contextual framework [33].

Another major subtheme involved evidence grounding through RAG. The reviewed studies consistently identified hallucination and factual inconsistency as major barriers to clinical adoption [18]. To address these limitations, multiple studies proposed integrating external biomedical literature, clinical guidelines, institutional protocols, and curated medical repositories into GenAI pipelines. RAG-based systems were frequently described as mechanisms for improving factual reliability, traceability, and evidence-based trust within clinical decision support environments [13, 16].

Thematic analysis also highlighted the increasing importance of domain-specific medical intelligence. Several studies emphasized the development of specialty-oriented foundation models and curated medical corpora designed for vertical clinical applications such as oncology, radiology, pathology, surgery, and critical care [14, 18, 21, 27, 36]. These systems improve contextual relevance and specialty-specific reasoning performance by embedding structured medical knowledge and domain-specific terminology into the model architecture.

Collectively, these findings suggest that perception and fact anchoring functions form the informational and cognitive foundation of GenAI clinical co-pilot systems. By improving contextual understanding, evidence retrieval, and multimodal interpretation, these capabilities support more reliable and clinically relevant AI-assisted healthcare environments.

D. Clinical Agency and Collaboration

The second major thematic domain identified in the findings involved clinical agency and collaboration, representing the operational layer of GenAI integration within healthcare workflows. This domain focuses on how GenAI systems support clinicians during routine clinical activities, documentation processes, communication tasks, and collaborative decision-making.

One of the most widely discussed applications involved ambient digital scribing and automated clinical documentation [12]. Several studies examined ambient AI systems capable of capturing physician-patient conversations and automatically generating structured clinical documentation, including SOAP notes, discharge summaries, and encounter records [7, 35]. These applications were consistently associated with reductions in administrative burden, documentation workload, and physician burnout [30].

Another major theme involved agentic clinical decision support (CDS). Unlike traditional rule-based CDS systems, GenAI systems increasingly function as interactive reasoning agents capable of differential diagnosis generation, triage support, contextual recommendation synthesis, and conversational clinical reasoning [15, 30, 54, 55]. Several studies described these systems as “clinical assistants,” “digital residents,” or “AI collaborators” designed to augment physician reasoning while preserving clinician supervision and final decision authority [18].

Personalized care and patient empowerment also emerged as important application areas. Multiple studies explored the integration of wearable devices, remote monitoring systems, and personalized patient care plans (PPCPs) into GenAI workflows [16, 28, 29]. These systems were designed to support individualized treatment coordination, shared decision-making, patient engagement, and long-term care continuity [10, 18, 25].

In addition, workflow optimization and interdisciplinary communication were repeatedly identified as key operational benefits [19]. GenAI systems were increasingly used to coordinate clinical tasks, streamline information retrieval, summarize large volumes of medical data, and facilitate communication across multidisciplinary healthcare teams [23]. These capabilities may help reduce cognitive overload and improve healthcare operational efficiency in increasingly complex clinical environments.

Overall, the findings indicate that GenAI is evolving beyond isolated documentation or chatbot functions toward broader collaborative clinical agency and collaboration integration. Within this emerging paradigm, AI systems increasingly function as workflow collaborators embedded within clinician-centered healthcare environments.

E. Governance and Responsibility

The third major thematic domain identified in the findings involved Governance and Responsibility, representing the institutional and ethical foundation required for safe GenAI deployment in healthcare settings. Across the reviewed

literature, governance mechanisms were consistently described as essential prerequisites for trustworthy clinical implementation.

One of the most prominent themes involved human-in-the-loop (HITL) oversight [34, 57]. Multiple studies emphasized that clinicians must remain the final reviewers, supervisors, and accountable decision-makers within AI-assisted healthcare systems [18, 32]. Rather than replacing physician expertise, GenAI systems were consistently positioned as supportive tools operating under clinician supervision [6, 41, 53]. This collaborative oversight structure was widely regarded as essential for maintaining patient safety and preserving clinical accountability [16].

Accountability and auditability also emerged as major governance concerns. Several studies discussed the need for traceable AI-assisted decisions, institutional governance committees, audit trails, and transparent documentation mechanisms capable of supporting legal and organizational accountability. Questions regarding liability attribution, clinician responsibility, and institutional oversight remain unresolved challenges in real-world healthcare deployment [9, 10, 32].

Privacy, security, and regulatory compliance represented another major subtheme. Because GenAI systems frequently process sensitive healthcare information, studies consistently emphasized compliance with GDPR, HIPAA, and healthcare cybersecurity standards. Several authors also highlighted risks related to data leakage, unauthorized access, and secondary use of patient information [8, 21, 42].

Ethics, fairness, and trustworthy AI were similarly recurring themes across the included studies. Multiple studies discussed algorithmic bias, demographic disparities, explainability limitations, and fairness concerns associated with large-scale GenAI deployment [9, 18]. Consequently, fairness audits, transparency mechanisms, explainable AI (XAI), and bias mitigation strategies were frequently proposed as critical components of responsible healthcare AI governance [8, 25, 43].

Collectively, the findings suggest that governance and responsibility mechanisms are not peripheral considerations, but foundational requirements for sustainable clinical AI integration. Successful implementation depends not only on technical performance, but also on the establishment of trustworthy human-AI collaboration frameworks supported by accountability, transparency, fairness, and clinician oversight.

IV. Discussion

A. Theoretical and Practical Contributions

This study makes several important theoretical and practical contributions to the emerging field of GenAI in healthcare. First, by systematically reviewing and synthesizing 41 studies published between 2023 and 2026, this research provides a comprehensive overview of the rapidly evolving literature on GenAI applications in healthcare. The findings consolidate fragmented evidence from medicine, healthcare informatics, artificial intelligence, and digital health into a coherent knowledge framework.

This study extends existing review literature that has primarily focused on ChatGPT performance, specialty-specific applications, or technical evaluations of LLMs [18, 47]. To the study's knowledge, this is among the first reviews to conceptualize GenAI as a clinical co-pilot through the integration of technological capabilities, clinical workflow support, and governance considerations. This broader perspective helps explain the evolution of GenAI from isolated tools toward collaborative clinical intelligence systems.

Second, this study proposes a conceptual framework that positions GenAI as a collaborative clinical co-pilot rather than a standalone decision-making system. The framework advances existing discussions beyond traditional clinical decision support systems by emphasizing the complementary relationship between human expertise and AI-enabled intelligence. This perspective contributes to the growing body of literature on human-AI collaboration and collaborative clinical intelligence.

Third, the thematic synthesis identifies three interrelated domains that characterize the emerging role of GenAI in healthcare: 1) Perception and Fact Anchoring, 2) Clinical Agency and Collaboration, and 3) Governance and Responsibility. These domains provide a structured theoretical foundation for understanding how GenAI systems acquire, interpret, and utilize clinical information while operating within healthcare workflows and regulatory environments.

From a practical perspective, the proposed framework may provide guidance for healthcare organizations, technology developers, and policymakers seeking to implement GenAI responsibly. The findings highlight the importance of evidence grounding, workflow integration, human oversight, transparency, and regulatory compliance as critical factors for successful deployment. Furthermore, the framework may provide a foundation for future empirical studies examining the effectiveness, safety, and long-term impact of GenAI clinical co-pilots in real-world healthcare settings.

B. From Automation to Collaboration

Traditional healthcare artificial intelligence systems primarily emphasized automation-oriented objectives, including disease prediction, risk stratification, image classification, and algorithmic screening [8, 17, 22]. Although many of these systems demonstrated strong benchmark performance, real-world clinical implementation remained limited due to poor workflow integration, insufficient interpretability, clinician distrust, and limited adaptability to complex healthcare environments [9, 18].

The findings suggest an emerging conceptual transition from automation-centered artificial intelligence toward collaboration-centered clinical intelligence. Unlike earlier predictive models that functioned primarily as isolated computational tools, contemporary GenAI systems increasingly operate as interactive clinical collaborators capable of contextual reasoning, multimodal information synthesis, evidence retrieval, and dynamic communication under clinician supervision [9, 35].

Within the proposed clinical co-pilot framework, the role of artificial intelligence shifts from replacing human decision-making toward augmenting clinician capabilities [35]. Rather than functioning as autonomous decision-makers, GenAI

systems support healthcare professionals by improving situational awareness, reducing cognitive burden, streamlining information retrieval, and facilitating workflow coordination [18, 35].

This transition is particularly important in healthcare because clinical decision-making inherently involves uncertainty, contextual judgment, ethical responsibility, interdisciplinary communication, and patient-centered reasoning. Traditional machine learning systems may generate accurate predictions, but they often lack the ability to explain reasoning pathways, contextualize recommendations, or engage in iterative clinical dialogue [16, 32]. In contrast, GenAI co-pilots demonstrate the ability to synthesize heterogeneous information sources, generate contextual explanations, and support collaborative reasoning processes within clinical environments.

Although the reviewed studies generally reported positive outcomes, the magnitude of these benefits varied across clinical applications. Administrative tasks, such as clinical documentation and ambient digital scribing, demonstrated more consistent performance improvements because they involve relatively standardized workflows and lower levels of clinical uncertainty. In contrast, applications involving diagnostic reasoning, treatment planning, and agentic clinical decision support exhibited greater variability, reflecting differences in domain complexity, multimodal data integration, evidence grounding capabilities, and the degree of required clinician oversight. These contextual differences suggest that the effectiveness of GenAI depends not only on model capability but also on the clinical context in which it is deployed.

Thematic synthesis further suggests that the future direction of healthcare AI may depend less on fully autonomous systems and more on collaborative intelligence architectures integrating human expertise with AI-supported reasoning[35]. Consequently, the clinical co-pilot paradigm may represent a more feasible, sustainable, and ethically responsible pathway for healthcare digital transformation.

C. Human-AI Partnership and Trust Calibration

The effectiveness of GenAI clinical co-pilots depends heavily on the establishment of appropriate human-AI partnerships and calibrated clinician trust. Across the reviewed literature, trust calibration emerged as a critical factor influencing healthcare AI adoption, workflow integration, and long-term clinical acceptance[35].

Overtrust and undertrust both represent substantial risks in AI-assisted healthcare environments [10]. Excessive reliance on AI-generated recommendations may increase vulnerability to hallucinations, factual inaccuracies, and automation bias [30]. Conversely, insufficient trust may prevent clinicians from utilizing potentially valuable decision support capabilities and operational efficiencies [37].

The reviewed studies consistently identified RAG, evidence grounding, and domain-specific medical intelligence as essential mechanisms for improving trust calibration [13, 51]. By anchoring outputs to authoritative medical literature, institutional guidelines, and curated biomedical knowledge sources, RAG-based systems improve factual consistency, transparency, and traceability. These mechanisms help transform GenAI from a probabilistic text generator into a more reliable evidence-supported clinical assistant [13].

Trust calibration is also associated with workflow integration quality. Studies examining ambient digital scribing and workflow-oriented clinical AI systems demonstrated that clinician acceptance increases when AI systems operate seamlessly within clinical workflows without introducing excessive operational disruption. Automated clinical documentation, structured note generation, and intelligent information summarization were associated with reduced administrative burden and lower physician burnout [12, 30].

Another important finding involved the emergence of agentic clinical decision support systems. Several studies described GenAI systems functioning as interactive reasoning agents capable of participating in diagnostic discussions, differential diagnosis generation, triage support, and multidisciplinary information synthesis [15, 55]. However, the reviewed literature consistently emphasized that these systems should remain clinician-supervised collaborative assistants rather than independent clinical authorities [28, 41].

Overall, the findings suggest that sustainable healthcare AI adoption depends not only on technical performance, but also on the development of trustworthy human-AI collaboration models characterized by transparency, explainability, contextual reliability, and workflow compatibility. Beyond technical performance, the heterogeneity observed across the reviewed studies suggests that successful implementation depends on organizational readiness, workflow redesign, clinician acceptance, and regulatory maturity. Consequently, similar GenAI technologies may produce different outcomes across healthcare settings, indicating that implementation context is as important as algorithmic performance.

D. Governance, Responsibility, and De-Identification Challenges

As GenAI systems transition from experimental research environments toward real-world healthcare implementation, governance and responsibility mechanisms become increasingly critical. The findings consistently suggest that technical performance alone is insufficient for sustainable healthcare AI deployment without corresponding institutional, ethical, and regulatory safeguards.

Human-in-the-loop oversight emerged as one of the most consistently emphasized governance principles across the reviewed studies [42, 56]. Regardless of the sophistication of GenAI systems, clinicians were repeatedly identified as the final reviewers, supervisors, and accountable decision-makers within healthcare environments [12, 28, 41]. This governance structure is particularly important because healthcare decisions involve substantial ethical responsibility, patient safety considerations, and legal accountability.

The reviewed literature also identified unresolved challenges regarding liability attribution and institutional accountability [10, 32]. Questions remain regarding how responsibility should be distributed when AI-assisted recommendations contribute to unsafe decisions, delayed diagnoses, or inappropriate treatments [8, 25, 41]. These concerns become increasingly complex as

GenAI systems evolve toward more autonomous agentic capabilities.

Privacy protection and healthcare data security were similarly identified as major implementation barriers. GenAI systems frequently process highly sensitive multimodal healthcare information, including electronic health records, radiological images, laboratory data, wearable sensor outputs, and free-text clinical narratives. Traditional de-identification approaches may become insufficient in large language model environments because contextual associations within unstructured clinical text may still permit indirect patient re-identification [17, 53].

Several studies emphasized that advanced LLMs possess strong contextual inference capabilities capable of reconstructing sensitive information from fragmented clinical details. Consequently, healthcare institutions may require more advanced privacy-preserving architectures, secure deployment infrastructures, and continuous monitoring systems to reduce data leakage risks [14, 52, 56].

The reviewed studies further emphasized the importance of transparency, fairness, explainability, and regulatory alignment. Bias inherited from historical healthcare datasets may contribute to unequal performance across demographic populations, clinical institutions, and healthcare settings. Therefore, fairness auditing, explainable AI mechanisms, and continuous governance evaluation were frequently proposed as essential components of trustworthy healthcare AI deployment [10, 32].

In addition, emerging regulatory frameworks such as the European Union AI Act and evolving FDA guidance may significantly influence future healthcare AI governance. However, current regulations continue to face challenges adapting to rapidly evolving GenAI systems characterized by dynamic reasoning capabilities, continuous model updates, and multimodal integration.

Notably, several reviewed studies also highlighted practical implementation challenges, including clinician skepticism, workflow disruption, and uncertainty regarding legal responsibility, hallucinations, and concerns over data privacy. These findings suggest that successful GenAI adoption depends not only on advances in model capability but also on organizational adaptation, effective governance, and sustained human oversight throughout clinical implementation.

Collectively, these findings suggest that governance and responsibility should not be viewed as secondary considerations, but rather as foundational infrastructure supporting safe, ethical, and sustainable healthcare AI integration.

E. Study Limitations

Several limitations should be acknowledged when interpreting these findings. First, only English-language publications were included in the analysis. Although English represents the dominant language of scientific communication, relevant studies published in other languages may have been excluded, potentially introducing language bias.

Second, the literature search was limited to three major databases: Web of Science, PubMed, and Scopus. While these databases provide extensive coverage of healthcare, medical informatics, and artificial intelligence research, relevant studies indexed in other databases or grey literature sources may not have been captured.

Third, this study employed a qualitative thematic synthesis approach rather than a quantitative meta-analysis. Given the heterogeneity of study designs, evaluation methods, and reported outcomes, direct statistical comparison across studies was not feasible. Consequently, the findings should be interpreted as conceptual and thematic insights rather than quantitative estimates of effectiveness.

Finally, GenAI is a rapidly evolving field characterized by continuous technological advancements and frequent model updates. As new foundation models, multimodal systems, and regulatory frameworks continue to emerge, some findings may require further validation and refinement over time. Therefore, the proposed framework should be viewed as an evolving conceptual foundation that may be expanded as the evidence base continues to grow.

F. Future Directions

The findings suggest several important future directions for GenAI clinical co-pilot research and healthcare implementation.

First, future studies should move beyond static benchmark evaluations toward longitudinal real-world clinical assessments. Many current studies rely heavily on benchmark datasets, examination-style evaluations, or proof-of-concept experiments. Although these approaches provide useful technical comparisons, they do not adequately capture long-term workflow integration, clinician adaptation, patient-centered outcomes, or operational sustainability [9, 36]. Future research should therefore prioritize longitudinal observational studies, implementation science approaches, and pragmatic clinical evaluations conducted within active healthcare environments.

Second, future clinical co-pilot systems will likely require deeper integration with multimodal healthcare ecosystems. The reviewed literature suggests that the next generation of healthcare AI may extend beyond isolated language processing toward comprehensive multimodal reasoning architectures integrating electronic health records, medical imaging, laboratory findings, wearable devices, physiological monitoring, and behavioral data streams. Such systems may facilitate more personalized, context-aware, and continuous healthcare support [8, 27].

Third, uncertainty quantification and explainability remain important future research priorities [58, 59]. Current GenAI systems frequently produce highly fluent outputs without explicitly communicating confidence levels or uncertainty boundaries [18, 36]. Future clinical co-pilots should incorporate transparent uncertainty signaling mechanisms capable of informing clinicians when recommendations require additional validation or specialist review. Such mechanisms may help reduce automation bias and improve clinical risk management.

Fourth, future governance frameworks must address distributed responsibility within collaborative human-AI healthcare systems. As GenAI systems become more deeply integrated into healthcare workflows, questions regarding legal liability, institutional accountability, and clinician responsibility

will become increasingly important [9, 10, 35]. Interdisciplinary collaboration among healthcare professionals, legal scholars, policymakers, and AI researchers will likely be necessary to establish operationally feasible governance standards.

Finally, the concept of clinical intelligence itself may continue to evolve. The reviewed literature suggests that healthcare AI is gradually transitioning from isolated decision support tools toward collaborative intelligence ecosystems integrating clinicians, patients, healthcare institutions, and AI-supported reasoning systems [28, 32]. Within this evolving paradigm, GenAI co-pilots may become foundational infrastructure supporting future human-centered digital healthcare environments.

V. Conclusion

This study systematically reviewed and synthesized recent research on GenAI in healthcare. Following the PRISMA 2020 framework, 41 studies published between 2023 and March 2026 were included in the final analysis. The findings indicate that GenAI is expanding beyond traditional decision-support functions and becoming more closely integrated into clinical workflows.

Three major themes emerged from the thematic synthesis: Perception and Fact Anchoring, Clinical Agency and Collaboration, and Governance and Responsibility. These themes reflect the key capabilities and challenges associated with the adoption of GenAI in healthcare. The findings show that technologies such as LLMs, multimodal foundation models, RAG, and agent-based systems are being applied across clinical documentation, decision support, patient communication, and care planning. At the same time, issues related to transparency, accountability, privacy protection, and human oversight remain central to successful implementation.

Based on these findings, this study proposes a Clinical Co-pilot Framework that integrates technological capabilities, clinical workflow support, and governance considerations. The framework may provide a conceptual perspective for understanding how GenAI can assist healthcare professionals while maintaining clinician responsibility for medical decision-making. The results suggest that the value of GenAI in healthcare depends not only on model performance but also on its ability to support clinical practice in a safe, reliable, and accountable manner. Effective deployment requires appropriate workflow integration, continuous evaluation, and clear governance mechanisms. Human oversight remains essential in ensuring the quality and safety of AI-assisted healthcare services.

Future research should focus on real-world implementation, long-term clinical impact, and the evolving relationship between healthcare professionals and AI systems. As GenAI technologies continue to develop. The future of healthcare AI may depend not on replacing clinicians, but on establishing effective human-AI partnerships that enhance clinical intelligence, support responsible decision-making, and improve patient-centered healthcare delivery.

References

[1] M. Elhaddad and S. Hamam, "AI-driven clinical decision support systems: an ongoing pursuit of potential," Cureus, vol. 16, no. 4, p. e57728, 2024.

[2] K. Singhal et al., "Large language models encode clinical knowledge," Nature, vol. 620, no. 7972, pp. 172-180, 2023.

[3] R. AlSaad et al., "Multimodal large language models in health care: applications, challenges, and future outlook," Journal of medical Internet research, vol. 26, p. e59505, 2024.

[4] Q. Jin et al., "Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine," NPJ Digital Medicine, vol. 7, no. 1, p. 190, 2024.

[5] G. Alain, J. Crick, E. Snead, C. C. Quatman-Yates, and C. E. Quatman, "Evaluating user interactions and adoption patterns of generative AI in health care occupations using claude: Cross-Sectional study," Journal of Medical Internet Research, vol. 27, p. e73918, 2025.

[6] S. Wojcik, A. Rulkiewicz, P. Pruszczyk, W. Lisik, M. Pobozy, and J. Domienik-Karlowicz, "Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era," Cardiol J, vol. 30, no. 6, pp. 1018-1025, 2023, doi: 10.5603/cj.97515.

[7] X. Chen, L. Zhou, J. Chen, G. Wang, and X. Li, "How is language intelligence evolving? A multi-dimensional survey of large language models," Expert Systems with Applications, vol. 304, 2026, doi: 10.1016/j.eswa.2025.130637.

[8] D. S. Comeau, D. S. Bitterman, and L. A. Celi, "Preventing unrestricted and unmonitored AI experimentation in healthcare through transparency and accountability," NPJ Digit Med, vol. 8, no. 1, p. 42, Jan 18 2025, doi: 10.1038/s41746-025-01443-2.

[9] Y. Artsi et al., "Challenges of Implementing LLMs in Clinical Practice: Perspectives," J Clin Med, vol. 14, no. 17, Sep 1 2025, doi: 10.3390/jcm14176169.

[10] T. Tung, S. M. N. Hasnaeen, and X. Zhao, "Ethical and practical challenges of generative AI in healthcare and proposed solutions: a survey," Front Digit Health, vol. 7, p. 1692517, 2025, doi: 10.3389/fdgth.2025.1692517.

[11] J. J. Andrew, J. Potier, N. Garcelon, A. Burgun, and M. Vincent, "Using large language models for temporal relation extraction from pediatric clinical reports," JAMIA open, vol. 8, no. 6, p. ooaf121, 2025, doi: 10.1093/jamiaopen/ooaf121 JAMIA.

[12] H. Wang et al., "An evaluation framework for ambient digital scribing tools in clinical applications," NPJ Digit Med, vol. 8, no. 1, p. 358, Jun 13 2025, doi: 10.1038/s41746-025-01622-1.

[13] B. B. Ozmen and P. Mathur, "Evidence-based artificial intelligence: Implementing retrieval-augmented generation models to enhance clinical decision support in plastic surgery," J Plast Reconstr Aesthet Surg, vol. 104, pp. 414-416, May 2025, doi: 10.1016/j.bjps.2025.03.053.

[14] J. Shen et al., "The foundational cornerstone for healthcare AI models: constructing multimodal corpora in the health sector," Chinese Science Bulletin, vol. 70, no. 26, pp. 4560-4568, 2025, doi: 10.1360/tb-2025-0185.

[15] N. Mehandru, B. Y. Miao, E. R. Almaraz, M. Sushil, A. J. Butte, and A. Alaa, "Evaluating large language models as agents in the clinic," NPJ Digit Med, vol. 7, no. 1, p. 84, Apr 3 2024, doi: 10.1038/s41746-024-01083-y.

[16] S. A. Rabbani et al., "Generative Artificial Intelligence in Healthcare: Applications, Implementation Challenges, and Future Directions," BioMedInformatics, vol. 5, no. 3, 2025, doi: 10.3390/biomedinformatics5030037.

[17] S. Guleria, J. Guptill, I. Kumar, M. McClintic, and J. C. Rojas, "Artificial intelligence integration in healthcare: perspectives and trends in a survey of U.S. health system leaders," BMC Digital Health, vol. 2, no. 1, 2024, doi: 10.1186/s44247-024-00135-3.

[18] S. Maity and M. J. Saikia, "Large Language Models in Healthcare and Medical Applications: A Review," Bioengineering (Basel), vol. 12, no. 6, Jun 10 2025, doi: 10.3390/bioengineering12060631.

[19] S. Frid et al., "Bridging generative AI and healthcare practice: insights from the GenAI Health Hackathon at Hospital Clinic de Barcelona," BMJ Health Care Inform, vol. 32, no. 1, Oct 15 2025, doi: 10.1136/bmjhci-2025-101640.

[20] H. Li, "AI Foundations in China's Medical Physiology Education: Pedagogical Practices and Systemic Challenges," Adv Med Educ Pract, vol. 16, pp. 1439-1453, 2025, doi: 10.2147/AMEP.S532951.

[21] C. Rainey, A. England, P. C. Murphy, A. A. Mohammad, Y. H. Hadi, and M. McEntee, "Large language models (LLMs) in radiography research: A narrative review," Radiography (Lond), vol. 32, no. 1, p. 103244, Jan 2026, doi: 10.1016/j.radi.2025.103244.

[22] R. Kawasaki, "How Can Artificial Intelligence Be Implemented Effectively in Diabetic Retinopathy Screening in Japan?," Medicina (Kaunas), vol. 60, no. 2, Jan 30 2024, doi: 10.3390/medicina60020243.

[23] J. D. Workum, D. van de Sande, D. Gommers, and M. E. van Genderen, "Bridging the gap: a practical step-by-step approach to warrant safe implementation of large language models in healthcare," Front Artif Intell, vol. 8, p. 1504805, 2025, doi: 10.3389/frai.2025.1504805.

[24] Y. Tsujimoto, M. Shiraishi, H. Yamanaka, H. Lee, M. Okazaki, and N. Morimoto, "Virtual patient modeling for generative-AI-assisted treatment decision-making in lymphedema care: AI tends to favor more aggressive treatment," Eur J Surg Oncol, vol. 52, no. 3, p. 111391, Mar 2026, doi: 10.1016/j.ejso.2026.111391.

[25] A. Pal et al., "Generative AI/LLMs for Plain Language Medical Information for Patients, Caregivers and General Public: Opportunities, Risks and Ethics," Patient Prefer Adherence, vol. 19, pp. 2227-2249, 2025, doi: 10.2147/PPA.S527922.

[26] O. G. Asai, N. Sabu, P. Gondode, S. Das, and G. Chauhan, "Comparative Evaluation of Popular Gen-AI Chatbots in Generating Patient Education Material on Pulmonary Artery Catheter Insertion," Ann Card Anaesth, vol. 29, no. 1, pp. 81-88, Jan 1 2026, doi: 10.4103/aca.aca_145_25.

[27] H. W. Jung, J. Y. Park, T. Holoubek, W. J. Kim, and J. Park, "Socially Assistive Robots in Mental Healthcare: Principles and Conceptual Framework for User-Centered Design," International Journal of Social Robotics, vol. 17, no. 11, pp. 2827-2851, 2025, doi: 10.1007/s12369-025-01323-5.

[28] R. J. Krumsvik and V. Slettvoll, "Artificial intelligence and health empowerment in rural communities and landslide- or avalanche-isolated contexts: real case at a fictitious location," Front Digit Health, vol. 7, p. 1655154, 2025, doi: 10.3389/fdgth.2025.1655154.

[29] M. M. Baig, C. Hobson, H. GholamHosseini, E. Ullah, and S. Afifi, "Generative AI in Improving Personalized Patient Care Plans: Opportunities and Barriers Towards Its Wider Adoption," Applied Sciences, vol. 14, no. 23, 2024, doi: 10.3390/app142310899.

[30] J. Pawelczyk, M. Kraus, S. Voigtlaender, S. Siebenlist, and M. C. Rupp, "Advancing Musculoskeletal Care Using AI and Digital Health Applications: A Review of Commercial Solutions," HSS J, vol. 21, no. 3, pp. 331-341, Aug 2025, doi: 10.1177/15563316251341321.

[31] J. L. Painter, D. Ramcharran, and A. Bate, "Perspective review: Will generative AI make common data models obsolete in future analyses of distributed data networks?," Ther Adv Drug Saf, vol. 16, p. 20420986251332743, 2025, doi: 10.1177/20420986251332743.

[32] P. Goktas and A. Grzybowski, "Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI," J Clin Med, vol. 14, no. 5, Feb 27 2025, doi: 10.3390/jcm14051605.

[33] F. Pesapane, R. Cuocolo, and F. Sardanelli, "The Picasso's skepticism on computer science and the dawn of generative AI: questions after the answers to keep "machines-in-the-loop"," Eur Radiol Exp, vol. 8, no. 1, p. 81, Jul 24 2024, doi: 10.1186/s41747-024-00485-7.

[34] J. S. Andersen and W. Maalej, "Design patterns for machine learning-based systems with humans in the loop," IEEE Software, vol. 41, no. 4, pp. 151-159, 2023.

[35] J. C. L. Chow and K. Li, "Large Language Models in Medical Chatbots: Opportunities, Challenges, and the Need to Address AI Risks," Information, vol. 16, no. 7, 2025, doi: 10.3390/info16070549.

[36] P. Gupta et al., "Rapid review: Growing usage of Multimodal Large Language Models in healthcare," J Biomed Inform, vol. 169, p. 104875, Sep 2025, doi: 10.1016/j.jbi.2025.104875.

[37] T. Mirzaei, L. Amini, and P. Esmaeilzadeh, "Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications," BMC Med Inform Decis Mak, vol. 24, no. 1, p. 250, Sep 9 2024, doi: 10.1186/s12911-024-02656-3.

[38] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," ACM Transactions on Information Systems, vol. 43, no. 2, pp. 1-55, 2025.

[39] Z. Zhu et al., "Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models," in Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 6748-6769.

[40] S. M. Sajjadi Mohammadabadi, B. C. Kara, C. Eyupoglu, C. Uzay, M. S. Tosun, and O. Karakuş, "A survey of large language models: evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications," Electronics, vol. 14, no. 18, 2025, doi: 10.3390/electronics.

[41] V. Chaudhuri, A. Brunelli, P. Tcherveniakov, and N. Chaudhuri, "Benchmarking Large Language Models Using a Best Evidence Topic Report in a Patient With Early Non-Small Cell Lung Cancer," Interdisciplinary CardioVascular and Thoracic Surgery, vol. 41, no. 2, p. ivag038, 2026, doi: 10.1093/icvts/ivag038 Interdisciplinary.

[42] R. Rokhshad et al., "The ethics and governance of large language models in dentistry: A framework for research and clinical implementation," J Dent, vol. 164, p. 106187, Jan 2026, doi: 10.1016/j.jdent.2025.106187.

[43] S. Pahune, Z. Akhtar, V. Mandapati, and K. Siddique, "The Importance of AI Data Governance in Large Language Models," Big Data and Cognitive Computing, vol. 9, no. 6, 2025, doi: 10.3390/bdcc9060147.

[44] P. D. Stetson et al., "Responsible Artificial Intelligence governance in oncology," NPJ Digit Med, vol. 8, no. 1, p. 407, Jul 4 2025, doi: 10.1038/s41746-025-01794-w.

[45] J. Rezwana and M. L. Maher, "Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems," ACM Transactions on Computer-Human Interaction, vol. 30, no. 5, pp. 1-28, 2023.

[46] A. B. Hendrickson, K. C. Gilkey, N. Stickler, and T. Taylor, "Evolving Healthcare Leadership in the Age of AI: A Narrative Review of Current Expected Transitions of Leadership in the era of Artificial Intelligence," International Journal of Scientific and Management Research, vol. 8, no. 8, pp. 17-29, 2025.

[47] B. G. Collaco et al., "The role of agentic artificial intelligence in healthcare: a scoping review," npj Digital Medicine, 2026.

[48] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, p. n71, Mar 29 2021, doi: 10.1136/bmj.n71.

[49] M. J. Page et al., "Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement," Journal of clinical epidemiology, vol. 134, pp. 103-112, 2021.

[50] B. Lim, I. Seth, M. Maxwell, R. Cuomo, R. J. Ross, and W. M. Rozen, "Evaluating the efficacy of large language models in generating medical documentation: a comparative study of ChatGPT-4, ChatGPT-4o, and Claude," Aesthetic Plastic Surgery, vol. 49, no. 20, pp. 5846-5857, 2025.

[51] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, "A Survey on RAG with LLMs," Procedia computer science, vol. 246, pp. 3781-3790, 2024.

[52] O. Shvetsova, D. Katalshov, and S.-K. Lee, "Innovative Guardrails for Generative AI: Designing an Intelligent Filter for Safe and Responsible LLM Deployment," Applied Sciences, vol. 15, no. 13, 2025, doi: 10.3390/app15137298.

[53] R. Dhawan, K. D. Brooks, O. Shauly, D. Shay, and A. Losken, "Ethical Considerations for Generative Artificial Intelligence in Plastic Surgery," Plast Reconstr Surg Glob Open, vol. 13, no. 6, p. e6825, Jun 2025, doi: 10.1097/GOX.0000000000006825.

[54] S. Bozyel et al., "Artificial intelligence-based clinical decision support systems in cardiovascular diseases," Anatolian journal of cardiology, vol. 28, no. 2, p. 74, 2024.

[55] L. Hughes, Y. K. Dwivedi, K. Li, M. Appanderanda, M. A. Al-Bashrawi, and I. Chae, "AI agents and agentic systems: Redefining global it management," vol. 28, ed: Taylor & Francis, 2025, pp. 175-185.

[56] S. Stamnes Karlsen, M. M. Yamin, E. Hashmi, B. Katt, and M. Ullah, "Securing large language models: A quantitative assurance framework approach," Journal of Information Security and Applications, vol. 97, 2026, doi: 10.1016/j.jisa.2025.104351.

[57] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," Artificial Intelligence Review, vol. 56, no. 4, pp. 3005-3054, 2023.

[58] B. Long, E. Liu, R. Qiu, and Y. Duan, "Explainable AI the latest advancements and new trends," arXiv preprint arXiv:2505.07005, 2025.

[59] M. Kirchhof, G. Kasneci, and E. Kasneci, "Position: Uncertainty quantification needs reassessment for large-language model agents," arXiv preprint arXiv:2505.22655, 2025.

APPENDIX

TABLE V. CHARACTERISTICS OF INCLUDED STUDIES

Author(s)	Main Technology	Study Type / Focus	Main Application	No
Andrew et al.	Llama 3 / Gemma / Mistral	Case Report	Extracting temporal relations from pediatric clinical reports to automate patient timelines for rare diseases.	[11]
Asai et al.	ChatGPT + Gemini	Original Article	Generating and validating patient education materials for pulmonary artery catheter insertion.	[26]
Baig et al.	GenAI + Wearables	Systematic review	Leveraging multimodal data for Personalized Patient Care Plans (PPCPs) and symbiotic intelligence.	[29]
Bozyel et al.	AI-based CDSS	Review / Perspective	Supporting cardiovascular disease diagnosis, risk assessment, and treatment optimization.	[54]
Chaudhuri et al.	GPT, Gemini, Co-pilot	Benchmarking	Evaluating LLM performance against "Best Evidence Topic" (BET) reports for early lung cancer decisions.	[41]
Chen et al.	LLM (Transformer)	Lifecycle survey	Mapping the evolution of language intelligence from word vectors to aligned multimodal agents.	[7]
Chow & Li	LLM (Transformer)	Review / Safety Analysis	Identifying opportunities and systemic risks in medical chatbots for patients and clinicians.	[35]
Comeau et al.	LLM + EHR	Governance framework	Ensuring transparency and accountability in EHR-based AI experimentation.	[8]
Dhawan et al.	GenAI + LLM	Ethical analysis	Exploring ethical considerations and administrative use cases specific to plastic surgery.	[53]
Frid et al.	LLM (Claude, Mistral, Titan)	Hackathon Analysis	Hospital-led prototyping of GenAI clinical tools under GDPR and EU AI Act compliance.	[19]
Goktas & Grzybowski	Symbolic AI + ML + Hybrid	Conceptual framework	Proposing a "Regulatory Genome" for trustworthy AI and adaptive oversight.	[32]
Guleria et al.	LLM / Generative AI	Executive survey	Analyzing perspectives and trends among US health system leaders on GenAI integration.	[17]
Gupta et al.	Multimodal LLM (MLLM)	Rapid Review	Mapping growing usage of multimodal data (imaging + text) in radiology, pathology, and ophthalmology.	[36]
Hughes et al.	AI Agents	Management Framework	To investigate the role of autonomous AI agents in advancing operational efficiency, intelligent decision-making, and collaborative IT management.	[55]
Jung et al.	Socially Assistive Robots	Conceptual Framework	User-centered design for assistive robots in mental healthcare utilizing multimodal evaluations.	[27]
Kawasaki	AI Screening	Implementation Study	Establishing a seven-step framework for effective diabetic retinopathy (DR) screening in Japan.	[22]
Khosravi et al.	LLM	Perspective / Survey	Identifying systemic challenges and diagnostic reasoning biases in clinical LLM adoption.	[9]
Krumsvik & Slettvoll	GPT-4 + OpenAI o3	Case Report / Rural	Implementing AI as a "virtual waiting room" for health empowerment in avalanche-isolated communities.	[28]
Li	AI + Machine Learning	Perspective / Education	Modernizing medical physiology education in China while addressing regional resource disparities.	[20]
Maity & Saikia	LLM (Various)	Systematic review	Comprehensive mapping of LLM capabilities in clinical decision support and medical education.	[18]
Mehandru et al.	LLM Agents + AI-SCE	Evaluation Framework	Transitioning assessment from static tests to high-fidelity "Objective Structured Clinical Examinations" (OSCE) for AI agents.	[15]
Mirzaei et al.	LLM + LDA Analysis	Thematic analysis	Synthesizing clinician voices on ethical implications, equity, and the patient-doctor dynamic.	[37]
Nazi & Peng	LLM + IGM	Comprehensive Review	Analyzing clinical documentation, diagnosis formulation, and evidence-based support systems.	[16]
Offodile et al.	GPT-4 + iLEAP	Institutional governance	Implementing the iLEAP lifecycle model for responsible AI deployment in oncology.	[44]
Ozmen & Mathur	RAG + LLM	Implementation study	Enhancing surgical clinical decision support by grounding responses in validated medical literature.	[13]
Pahune et al.	LLM Data Governance	Review / Framework	Establishing a five-pillar data governance model focusing on compliance and bias mitigation.	[43]
Painter et al.	LLM + Knowledge Graphs	Perspective Review	Exploring the potential for GenAI to automate data standardization and code conversion in healthcare.	[31]

Pal et al.	LLM	Review / Ethics	Evaluating plain language medical information for non-specialists and its impact on health literacy.	[25]
Pawelczyk et al.	AI + Digital Health	Commercial Review	Evaluating commercial AI solutions for musculoskeletal (MSK) triage, diagnostics, and surgical planning.	[30]
Pesapane et al.	GenAI + Radiomics	Narrative review	Discussing the "machines-in-the-loop" paradigm to enhance clinical utility in medical imaging.	[33]
Rainey et al.	LLM (Various)	Narrative Review	Automating radiography research stages, report structuring, and trial participant recruitment.	[21]
Rokhshad et al.	LLM	e-Delphi consensus	Creating an ethical governance framework for LLM implementation in clinical dentistry.	[42]
Sajjadi Mohammadabadi et al.	LLM Evolution	Survey	LLMs are increasingly adopted across diverse industries to support intelligent content generation, decision-making, knowledge management, and process automation.	[40]
Shen et al.	Multimodal Corpora	Framework / Corpus	Proposing "Disease-Scenario Association Matrices" as the cornerstone for healthcare vertical large models.	[14]
Shvetsova et al.	AVI Filter + RAG	Design / PoC Study	Implementing intelligent guardrails to filter toxicity and verify facts in responsible LLM deployment.	[52]
Stamnes Karlsen et al.	Quantitative Assurance	Methodology Study	Developing a framework to secure LLMs against prompt injection and toxic outputs in clinical apps.	[56]
Tsujimoto et al.	ChatGPT-4o	Virtual Patient Modeling	Identifying AI tendencies toward aggressive treatment in assisted decision-making for lymphedema care.	[24]
Tung et al.	GenAI	Systematic Review	Analyzing systemic bias, liability, and data privacy risks with proposed socio-technical solutions.	[10]
Wang et al.	Ambient AI + SCRIBE	Evaluation framework	Standardizing assessment of Ambient Digital Scribing (ADS) tools for clinical documentation.	[12]
Wójcik et al.	GPT-4	Invited Review	Mapping GPT-4's added value in clinical notes, triage, and cardiology consultations.	[6]
Workum et al.	LLM + ACUTE	Implementation roadmap	Providing a step-by-step approach for safe LLM implementation using the ACUTE mnemonic.	[23]