Sci

MDPI

Review

Applications and Challenges of Retrieval-Augmented Generation (RAG) in Maternal Health: A Multi-Axial Review of the State of the Art in Biomedical QA with LLMs

Adriana Noguera [1], Andrés L. Mogollón-Benavides [1], Manuel D. Niño-Mojica [1], Santiago Rua [1], Daniel Sanin-Villa [2] and Juan C. Tejada [3,4,*]

[1] ECBTI, Universidad Nacional Abierta y A Distancia, Bogota 111511, Colombia; adriana.noguera@unad.edu.co (A.N.); almogollonb@unadvirtual.edu.co (A.L.M.-B.); mdninom@unadvirtual.edu.co (M.D.N.-M.); santiago.rua@unad.edu.co (S.R.)
[2] Área de Industria, Materiales y Energía, Universidad EAFIT, Medellín 050022, Colombia; dsaninv2@eafit.edu.co
[3] Artificial Intelligence and Robotics Research Group (IAR), Universidad EIA, Envigado 055428, Colombia
[4] Department of Engineering Studies for Innovation, Universidad Iberoamericana Ciudad de México, Prolongación Paseo de la Reforma 880, Colonia Lomas de Santa Fé, Ciudad de México 01219, Mexico
* Correspondence: juan.tejada@ibero.mx

Abstract

The emergence of large language models (LLMs) has redefined the potential of artificial intelligence in clinical domains. In this context, retrieval-augmented generation (RAG) systems provide a promising approach to enhance traceability, timeliness, and accuracy in tasks such as biomedical question answering (QA). This article presents a narrative and thematic review of the evolution of these technologies in maternal health, structured across five axes: technical foundations of RAG, advancements in biomedical LLMs, conversational agents in healthcare, clinical validation frameworks, and specific applications in obstetric telehealth. Through a systematic search in scientific databases covering the period from 2022 to 2025, 148 relevant studies were identified. Notable developments include architectures such as BiomedRAG and MedGraphRAG, which integrate semantic retrieval with controlled generation, achieving up to 18% improvement in accuracy compared to pure generative models. The review also highlights domain-specific models like PMC-LLaMA and Med-PaLM 2, while addressing persistent challenges in bias mitigation, hallucination reduction, and clinical validation. In the maternal care context, the review outlines applications in prenatal monitoring, the automatic generation of clinically validated QA pairs, and low-resource deployment using techniques such as QLoRA. The article concludes with a proposed research agenda emphasizing federated evaluation, participatory co-design with patients and healthcare professionals, and the ethical design of adaptable systems for diverse clinical settings.

Keywords: retrieval-augmented generation; biomedical LLMs; maternal health; clinical validation; telemedicine

check for updates

Academic Editors: Haridimos Kondylakis and Michael Sheng

Received: 4 August 2025
Revised: 27 August 2025
Accepted: 9 October 2025
Published: 16 October 2025

Citation: Noguera, A.; Mogollón-Benavides, A.L.; Niño-Mojica, M.D.; Rua, S.; Sanin-Villa, D.; Tejada, J.C. Applications and Challenges of Retrieval-Augmented Generation (RAG) in Maternal Health: A Multi-Axial Review of the State of the Art in Biomedical QA with LLMs. *Sci* **2025**, *7*, 148. https://doi.org/10.3390/sci7040148

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Generative artificial intelligence (AI) has shown great potential for facilitating access to medical information, but its direct adoption in obstetric health remains problematic due to the phenomenon of hallucinations, in which incorrect or unverified responses are generated. To mitigate this risk, the retrieval-augmented generation (RAG) approach

has been proposed as a more reliable method, as it integrates validated evidence, such as the scientific literature or clinical guidelines, in real time. For example, Alkhalaf et al. [1] demonstrated that RAG significantly improves accuracy when extracting clinical information from electronic medical records. Similarly, a recent study evaluated clinical workflows incorporating RAG in 2000 real cases, showing substantial improvements in clinical decision-making [2].

A systematic review published in 2025 further highlights the growing adoption of RAG in healthcare, while identifying persistent gaps in ethics and evaluation metrics within these systems [3]. In specific applications, Ozmen et al. [4] applied RAG in plastic surgery, improving both the accuracy and transparency of AI-generated clinical recommendations. Moreover, Ji et al. [5] emphasized the need to investigate demographic bias in RAG systems, underscoring the urgency of developing more equitable approaches.

In recent years, interest in biomedical language models and retrieval-based architectures has expanded exponentially, as evidenced by a sustained increase in publications and clinical applications reported in recent bibliometric analyses [6]. This growth has been accompanied by the emergence of specialized resources such as the OxMat dataset, which consolidates a large volume of multimodal maternal and perinatal health records, thereby strengthening the representativeness of diverse populations in the training and validation of AI systems [7].

This article provides a critical review of the use of RAG in maternal health, structured around five key themes: technical foundations, recent biomedical models, agents in obstetric telemedicine, metrics and clinical safety, and specific applications in prenatal and perinatal care. The objective is to rigorously evaluate the benefits and limitations of these technologies in a high-risk environment, with particular attention given to safety, ethics, and equity.

The novelty of this review lies in its comprehensive and critical approach to maternal health and obstetric telemedicine, integrating the most recent literature on biomedical language models and RAG systems. Unlike previous reviews, it assesses not only technical advances but also clinical applicability, the safety of recommendations, and the representativeness of the datasets employed. The synthesis indicates that optimized RAG implementations, such as MedRAG and those evaluated in recent benchmarks [8], can improve the accuracy of general and specialized LLMs in medical question-answering tasks by up to 18%, thereby enhancing factual reliability and sensitivity in the detection of obstetric warning signs. In addition, this review identifies persistent gaps in clinical validation, equity, and data diversity, providing clear guidance for future research and for the safe application of RAG systems in high-risk settings.

1.1. Background

Generative artificial intelligence (generative AI) encompasses a set of machine learning techniques that enable systems to produce new, coherent content by learning patterns from large-scale data collections [9]. In recent years, this branch of AI has advanced rapidly, particularly with the emergence of LLMs. A key catalyst for this progress was the COVID-19 pandemic, which abruptly reshaped the global economy, social interactions, and most notably the healthcare sector. As the first line of response, healthcare systems faced an unprecedented overload, prompting an urgent demand for efficient digital solutions.

The pandemic acted as a turning point that accelerated digital transformation across many domains (healthcare most of all) [10]. Overburdened medical facilities and the need to preserve physical distancing encouraged widespread adoption of technology-based alternatives [11]. Within this setting, telemedicine gained renewed prominence. Although some institutions had already implemented remote-consultation services, the public-health

emergency promoted their broad uptake as a means of maintaining access to medical care [12]. This evolution also eased the incorporation of emerging technologies such as generative AI into clinical practice, fostering the development of conversational agents, decision-support tools, and medical-recommendation systems [10].

In recent years, and especially in the post-pandemic era, remarkable progress has been reported in what many authors term the fifth industrial revolution. This emerging paradigm is defined by the convergence of digital, physical, and biological technologies, in which machine learning, deep learning, chatbot, and artificial intelligence applications occupy a central position. The impact has been so evident that leading corporations now devote a substantial share of their resources to technological solutions designed to automate processes across multiple sectors, including healthcare [9,13].

In parallel, other studies underscore the same trend, highlighting that the confluence of these technologies is reshaping organisational strategies on a global scale [14]. The magnitude of this transformation has prompted large enterprises to increase their investment in automation and intelligent systems that can streamline workflows in areas such as diagnostics, logistics, and patient management within the healthcare sector [15].

Maternal health, in particular, represents a domain in which reliable technological tools can make a measurable difference by improving care delivery, reducing risks, and ensuring secure and useful information exchange. Mobile applications, digital platforms, and web-based solutions offer considerable potential to enhance the quality, accessibility, and safety of maternal-care services [16]. Their effectiveness, however, depends on thoughtful design, rigorous validation procedures, and professional oversight that safeguard reliability and applicability in real clinical settings [17].

Within this landscape, generative artificial intelligence has begun to emerge as a strategic instrument for strengthening maternal care in telemedicine environments. Recent work demonstrates that such technologies can process large volumes of clinical data, produce automated responses, and deliver personalised recommendations [9,15]. For example, LLMs can be embedded in virtual assistants that guide pregnant users regarding warning signs, key dates, or birth preparation, thereby lowering risks and enhancing care experiences, especially in communities with limited access to specialised services [18]. Nevertheless, deployment of these systems in sensitive clinical scenarios demands robust ethical frameworks, continuous medical supervision, and technical validation to guarantee the accuracy, safety, and relevance of the interactions they generate [9].

1.2. Large Language Models (LLMs)

LLMs are artificial intelligence systems that are built on deep-neural-network architectures, most notably transformers, that are trained on massive text corpora to understand, generate, and manipulate natural language at an advanced level. These models have transformed natural language processing, enabling tasks such as automatic translation, text summarisation, information extraction, and question answering. Their main strength lies in learning deep contextual representations of language, which allows adaptation to diverse domains and task-specific requirements [19].

LLM development has been marked by increasingly specialised architectures. Among the most prominent is the Generative Pre-trained Transformer (GPT) family, which has shown excellent performance in cancer phenotype extraction and clinical diagnosis tasks, outperforming earlier models and rule-based approaches. GPT-4, for example, delivers notable gains in precision and recall when identifying complex diagnoses within electronic medical records [20,21].

Other noteworthy models include the PaLM (Pathways Language Model), featuring a larger parameter scale and enhanced reasoning abilities, and BioGPT, which is tailored

to the biomedical domain through additional training on the scientific literature, enabling high-accuracy extraction of drug–disease relationships [22]. MedPaLM pushes the field toward multimodal integration by combining textual data with medical images to answer expert-level clinical questions [23].

Beyond core language processing, LLMs have proved valuable for improving equity and generalisation in automated medical-diagnosis systems. Generative models can create synthetic data that mitigate demographic bias and strengthen classifier robustness in out-of-distribution scenarios, as demonstrated in digital= histopathology and chest radiology studies [24]. Nonetheless, practical deployment poses challenges: updating a model's knowledge base demands costly retraining, and there is a risk of perpetuating or amplifying biases present in the source data [24].

Systematic evaluation of LLMs in healthcare applications underscores the need for diverse, representative validation datasets and metrics to ensure fairness and clinical effectiveness. The recent literature stresses that LLM performance varies with diagnostic modality and training-data quality, so clinical integration should proceed under rigorous validation protocols [25].

1.3. Retrieval-Augmented Generation (RAG)

Retrieval-augmented generation (RAG) is a hybrid approach that integrates the generative capabilities of large language models (LLMs) with information retrieval modules, allowing dynamic access to external and up-to-date knowledge bases during text generation [1,26].

A typical RAG architecture includes (i) a semantic-retrieval component that identifies and selects the most relevant documents or passages in response to a query and (ii) a generative model that synthesises a coherent answer using the retrieved material [26–28]. This integration produces responses that are contextually relevant and traceable because the specific sources underpinning each statement can be reported. Key advantages over stand-alone generative models include continuously updatable knowledge, improved factual accuracy, and a reduced risk of hallucinations or errors caused by outdated information [26–28].

Recent studies indicate that RAG adoption in medicine improves the quality and personalisation of clinical recommendations and enhances efficiency in extracting and synthesising information from electronic medical records [26,27]. For instance, web-based medical services equipped with RAG deliver more accurate, context-aware recommendations, while RAG-driven summarisation of clinical notes facilitates the identification of risk factors and the extraction of key details for decision-making [1,26].

Nevertheless, implementation challenges persist, including the need for efficient retrieval engines, continually updated databases, and clinical validation of generated answers. The literature suggests that RAG success depends heavily on the quality and diversity of the underlying information sources and on the model's ability to interpret and synthesise retrieved evidence accurately [27–30]. Despite these hurdles, current evidence supports RAG's potential to address limitations of purely generative models and move towards more reliable, practice-ready AI systems in healthcare.

Importantly, RAG does not completely eliminate hallucinations; however, it does help reduce their frequency and severity, particularly when combined with specialized biomedical models such as BioGPT or Med-PaLM, rather than general-purpose LLMs that often introduce biases and clinical errors [2,31]. The integration of RAG with these biomedical models represents a safer approach for supporting obstetric care in telemedicine settings.

Regarding evaluation, it is insufficient to rely exclusively on traditional linguistic metrics such as BLEU or ROUGE [32]. In high-risk domains such as maternal health, it is necessary to incorporate clinical metrics that reflect safety and relevance. These

include (i) sensitivity in detecting obstetric warning signs (e.g., recall for preeclampsia, hemorrhage, or convulsions), (ii) the critical error rate, which quantifies potentially harmful recommendations, and (iii) the system's ability to prioritize high-risk pregnant women in triage [5]. Such metrics enable evaluation not only of linguistic consistency but also of the clinical utility and safety of the systems.

The adoption of biomedical models also faces structural barriers, including a lack of comprehensive clinical validation, limited interoperability, and the absence of clear regulatory frameworks—factors that have been widely documented in recent reviews [33]. Moreover, recent studies show that general-purpose models can, in some cases, match or even outperform pre-trained biomedical models in previously unseen clinical tasks, reinforcing the advantage of architectures such as RAG, which leverage up-to-date and verifiable external information rather than relying solely on static training data [34].

1.4. Biomedical Question Answering (QA)

Biomedical question answering (QA) denotes the ability of artificial intelligence systems, particularly LLMs, to provide automatic, accurate, and context-aware answers to natural language questions about medical and biomedical topics. Such systems are designed to interpret complex queries, identify relevant information across vast collections of the scientific literature, clinical guidelines, and medical databases, and generate clear, evidence-based responses. The primary goal is to give health professionals and patients efficient, direct access to reliable, up-to-date medical knowledge that can inform clinical decision-making [23].

The clinical significance of biomedical QA lies in its capacity to streamline the retrieval and synthesis of pertinent information, ease cognitive load, and support real-time decisions. According to [23], QA systems powered by LLMs achieve performance that is comparable to that of human specialists when addressing complex clinical questions, a capability that is particularly valuable in scenarios where rapid, precise access to scientific evidence can influence patient safety and care quality. Moreover, integrating biomedical QA into clinical platforms helps lower error rates, improves service quality, and fosters personalised care by returning traceable answers that are grounded in recent evidence. Consequently, biomedical QA serves as a strategic tool for continuous professional development and wider access to expert medical knowledge [23].

1.5. Maternal Health in Telemedicine

Maternal health within the context of telemedicine has gained increasing importance in recent years, as it facilitates access to quality medical services for pregnant women regardless of their geographic location or socioeconomic status. Various studies conducted in Latin America have shown that implementing telemedicine for pregnant individuals improves access to medical care, increases the awareness of warning signs during pregnancy, childbirth, and the postpartum period, and enhances gestational monitoring. For instance, virtual educational interventions have helped pregnant women better recognise danger signs, thereby improving their ability to seek timely care and prevent severe complications [35]. Additionally, telemedicine has enabled the continuity of prenatal care during public health emergencies, such as the COVID-19 pandemic, helping to reduce logistical barriers and strengthen remote doctor–patient relationships [35,36].

Among the primary benefits of telemedicine in maternal health are remote education, telemonitoring, and the collection of clinical data at a distance. Remote education, delivered via teleconsultations and tele-guidance, has proven to be effective in increasing pregnant individuals' knowledge about disease prevention, such as iron-deficiency anemia, and in managing risk conditions [35]. Telemonitoring, in turn, enables proactive and personalised

patient follow-up, optimising medical resources and improving the quality of prenatal care, which ultimately leads to better perinatal outcomes [35]. Likewise, remote data collection supports epidemiological surveillance and maternal health research by identifying trends and assessing the real-time impact of interventions [35,37].

However, telemedicine also faces challenges, including the need to ensure equitable access to technology, protect data confidentiality, and provide training for healthcare professionals. Despite these limitations, the scientific literature consistently supports telemedicine as a viable and effective strategy for improving maternal health, especially in high-risk populations or those with limited access to in-person services. Accordingly, the development of programmes and public policies that promote its use, along with regulatory frameworks and digital literacy strategies, is recommended to maximise benefits and minimise associated risks [35–38].

2. Review Methodology

Given the rapid growth of the literature on the use of language models and retrieval-augmented generation (RAG) techniques in clinical contexts, it is essential to establish a systematic search and selection framework that guarantees both transparency and rigor. This section outlines the databases consulted, the inclusion and exclusion criteria applied, the article selection process, and the organization of the retrieved information into thematic areas. These steps ensure that the findings presented in the following sections provide an updated, critical, and methodologically robust overview.

The review methodology adopted in this work is designed to provide a comprehensive and rigorous examination of the literature on biomedical models and retrieval-augmented architectures. Since one of the major challenges identified in the field is the lack of diversity in clinical datasets [6], particular attention was given to integrating sources that increase representativeness. Among these, the OxMat dataset stands out as a benchmark resource for research on pregnancies and births across diverse clinical contexts, strengthening the relevance and applicability of the evidence reviewed [7].

This review combines a narrative approach with a structured critical reflection aimed at designing a text-mining pipeline that is capable of generating clinically valid QA pairs for training intelligent agents in maternal health. Given the multidisciplinary nature of the problem, which intersects natural language processing, generative artificial intelligence, clinical validation, and maternal health, a flexible yet rigorous methodological strategy has been adopted. The overall methodological framework is summarised in Figure 1, which outlines the main stages of the review, from thematic definition and the literature retrieval to synthesis and analysis by axis.

2.1. Thematic Axes of the Review

The reviewed literature is organised into five thematic axes, defined by their technical and clinical relevance to the development of the proposed system. This categorisation emerged from a functional analysis of the key components of the pipeline and allows the state of the art to be structured in direct alignment with the project's objectives. The selected axes and their corresponding information search criteria are detailed below:

- Fundamentals of RAG in biomedical QA systems: architecture, retrieval mechanisms, and traceability;
- Biomedical LLMs and clinical QA generation: performance, training domains, and applicability;
- Use of QA as input for medical conversational agents: datasets, fine-tuning, and health-related interaction;

- Clinical validation and explainability of QA systems: clinical concordance, traceability, and transparency;
- Specific applications of QA and NLP in maternal health: tools, use cases, and population coverage.

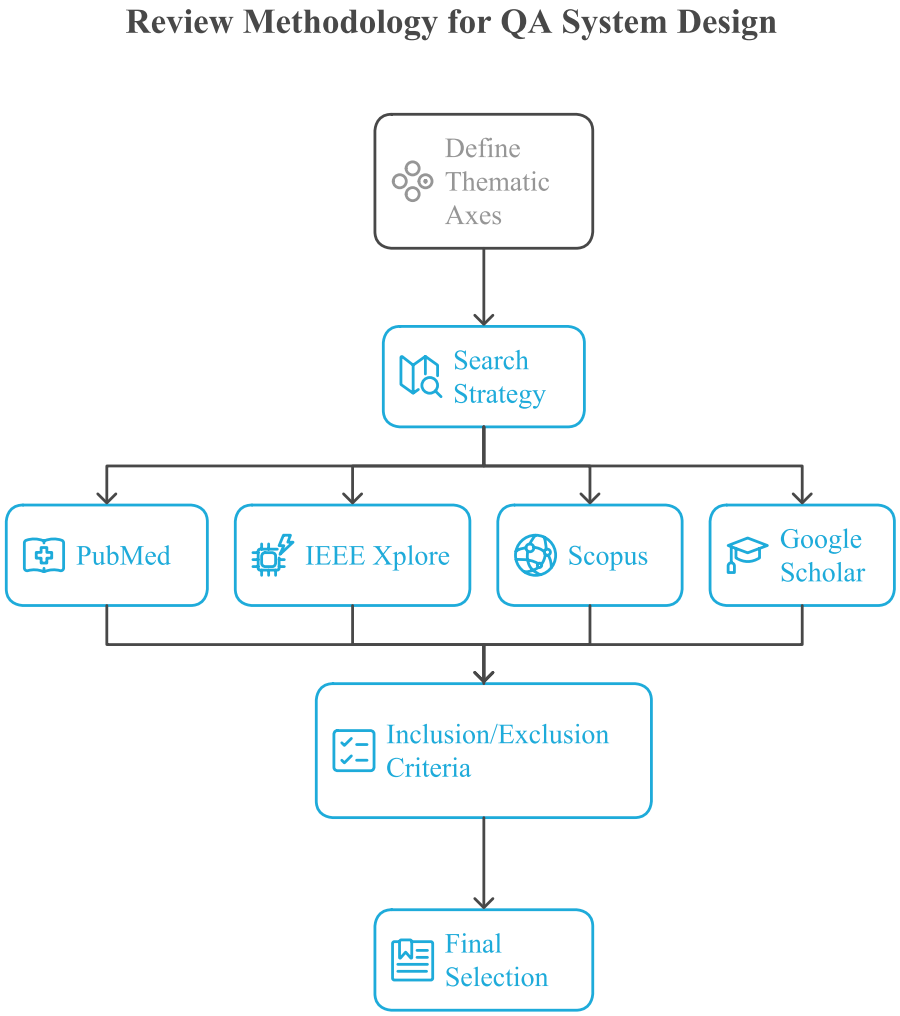

Figure 1. Methodological framework for the review.

2.2. Search Strategy and Sources

The literature search was conducted across PubMed, IEEE Xplore, Scopus, and Google Scholar. Although some articles were hosted on arXiv, these were retrieved via Google Scholar rather than through direct search on the arXiv platform. The search period spanned from 2022 to 2025. Search queries were adapted to the syntax and requirements of each engine and organised according to the five thematic axes previously defined.

As summarized in Table 1, the search strategies were designed to ensure comprehensive coverage of relevant biomedical and clinical domains.

For the evaluation of the models, particular emphasis was placed on metrics that capture both factual accuracy and the clinical safety of the generated responses. These include precision, sensitivity in detecting obstetric warning signs, and the critical error rate, all of which have been widely adopted in the recent literature for the validation of LLMs in medical contexts [39]. These metrics were preferred over alternatives that are focused solely on linguistic performance, as they better reflect dimensions of safety and clinical applicability, consistent with current recommendations [40].

Table 1. Thematic axes, databases, and search equations.

Thematic Axis	Database	Search Equation
Axis 1: Fundamentals of RAG in Medicine	PubMed	(‘‘retrieval augmented generation’’[Title/Abstract] OR RAG[Title/Abstract]) AND (medicine OR biomedical)
Axis 1: Fundamentals of RAG in Medicine	IEEE Xplore	‘‘retrieval augmented generation’’ AND (medical OR clinical)
Axis 2: Biomedical LLMs and QA Generation	Scopus	TITLE-ABS-KEY(‘‘biomedical language model’’ OR ‘‘BioGPT’’) AND ‘‘question answering’’ AND PUBYEAR > 2022
Axis 2: Biomedical LLMs and QA Generation	PubMed	(‘‘biomedical language model’’ OR ‘‘Med-PaLM’’ OR ‘‘BioGPT’’) AND (‘‘question answering’’)
Axis 3: QA as Input for Intelligent Agents	PubMed	(‘‘question answering’’ AND ‘‘chatbot’’ OR ‘‘conversational agent’’) AND (healthcare OR telemedicine)
Axis 4: Clinical Validation and Explainability	Scopus	TITLE-ABS-KEY(‘‘clinical validation’’ OR ‘‘traceability’’ OR ‘‘explainable AI’’) AND (‘‘medical question answering’’) AND PUBYEAR > 2022
Axis 5: Applications in Maternal Health	Google Scholar	‘‘maternal health QA system’’ OR ‘‘telemedicine agents pregnancy care’’ AND 2023..2025

With respect to retrieval methods, priority was given to RAG architectures due to their ability to integrate updated and verifiable evidence in real time, offering a clear advantage over traditional approaches such as BM25 or Dense Passage Retrieval. This decision is supported by recent systematic reviews demonstrating the superior performance of RAG in the extraction and synthesis of biomedical information [3], as well as comparative studies with pre-trained biomedical models, which reveal that, without the integration of external knowledge, such models may face limitations when applied to previously unseen clinical scenarios [34].

2.3. Inclusion and Exclusion Criteria

Articles were selected based on thematic alignment with the defined axes, recency (i.e., publications from 2022 to 2025), and academic visibility, assessed through citation count or recurrence across search results. The review included studies from recognised academic sources and relevant preprints, prioritising those with greater technical and clinical contributions to the design of question–answer generation systems in healthcare. Inclusion was determined by the relevance of content to the project context rather than by formal editorial criteria.

To complement the thematic analysis of the reviewed literature, a word cloud was generated. This visualization highlights the most frequent technical concepts, emphasizing the centrality of notions such as retrieval-augmented generation (RAG), large language models (LLMs), maternal health, telemedicine, and clinical validation, among others. Figure 2 graphically summarizes the relative prominence of these terms and supports the coherence of the multi-axial framework proposed in this review.

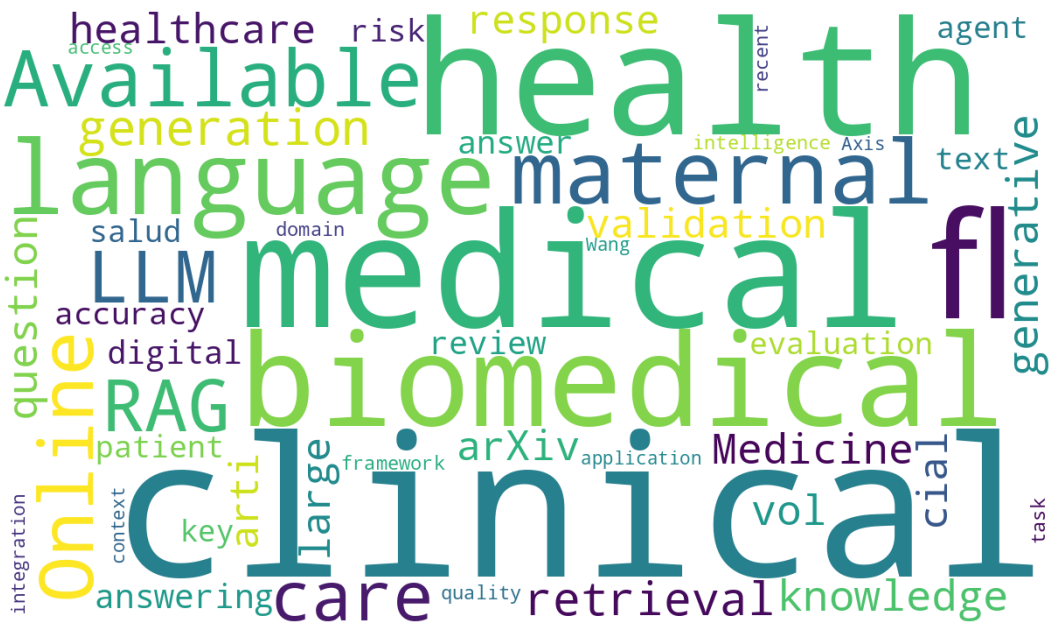

Figure 2. Conceptual synthesis from literature review.

3. State-of-the-Art Development by Thematic Axes

Generative artificial intelligence has transformed the way unstructured data(such as text, voice, audio, and video) are processed and analysed, with key applications emerging in domains like medicine. Tasks that once depended exclusively on human expertise, such as clinical diagnosis, knowledge generation, or medical question answering, can now be automated with increasing levels of accuracy [41].

Within the scope of this literature review, the state of the art is structured around five thematic axes, corresponding to the technical dimensions of the selected problem domain. On the one hand, understanding how RAG models operate in telemedicine and which LLMs are best suited for biomedical contexts helps to ground the proposed system's architecture. On the other hand, exploring how QA is used to train conversational agents, together with the requirements of traceability and clinical validation, is essential to ensure that the generated responses are trustworthy in telemedicine environments. Finally, given the specific focus of this review on maternal health monitoring, it is essential to assess whether prior research has effectively applied these technologies to this particular domain.

At the intersection of artificial intelligence and medicine, advances in question–answering systems are transforming access to and the use of clinical knowledge. Within this emerging field, five key axes have been identified that represent the development and application of these technologies, each with significant implications for medical practice and health research.

The first axis focuses on the evolution of RAG systems and their advanced versions, which enhance the ability of language models to generate accurate and context-aware responses in biomedical environments [42]. These systems combine information retrieval with text generation, thereby improving the precision of medical applications and demonstrating that AI systems can enhance both clinical information retrieval and the generation of contextually relevant answers [43].

The second axis highlights the development of LLMs that are trained in telemedicine domains. These specialised models are fine-tuned on medical terminology, clinical studies, and the scientific literature, optimising interactions between health professionals and automated information systems. Evidence shows that generative models have improved the interpretation of medical texts, aiding clinical decision-making and the automation of key healthcare processes [44].

The third axis explores the use of QA systems as foundational inputs for clinical conversational agents. These tools facilitate communication between professionals and patients by accurately interpreting and answering medical questions, thereby improving healthcare accessibility and efficiency. At the same time, they help assess the impact of conversational agents in private medical practice, providing more effective communication pathways and access to specialised information [45].

The fourth axis addresses clinical validation and explainability in QA systems, focusing on rigorous methods to evaluate the reliability of generated responses. These processes aim to ensure transparency in decision-making, which is an essential factor for building trust and achieving adoption in medical environments. The literature also identifies standards for quality validation in specialised health centres, underlining the importance of performance assessment for AI systems in medicine [46].

Finally, the fifth axis analyses the specific application of these technologies in maternal health, a field of innovation with direct impact on the well-being of pregnant women. The Pan American Health Organization (PAHO) [47] highlights how digital health technologies have supported maternal condition monitoring and diagnostics across Latin America, improving prenatal and neonatal care through advanced digital tools.

These five axes not only reflect the progress of artificial intelligence in medicine but also outline a future in which interactions between healthcare professionals and automated systems become increasingly seamless, trustworthy, and enriching. The following subsections detail the most relevant findings from the literature review according to the five thematic axes.

3.1. Axis 1: RAG and Advanced RAG in Biomedical QA Systems

The implementation of LLMs in biomedicine presents considerable challenges due to the sensitive and critical nature of medical information. A major limitation is the generation of incorrect responses, technically referred to as hallucinations, which may result from biased or outdated training data, limitations that are inherent in the model's prediction mechanism, the misinterpretation of the user's input, or poorly formulated queries [48,49]. In this context, retrieval-augmented generation (RAG) emerges as a promising solution, enabling LLMs to retrieve relevant external information by combining text generation with semantic search techniques for specialised content access [32]. In maternal health, this translates into the accurate retrieval of relevant passages from clinical guidelines using vector-based representations that enhance information relevance. Moreover, each response can be enriched with metadata such as the source, page, or originating document, supporting cross-verification and traceability by clinical staff [50].

Recent studies have reported the development of various frameworks and RAG-based systems, exploring different configurations and optimisations. For instance, ref. [41] presents Ascle, a clinical-text-generation tool for several NLP tasks, including QA, summarisation, translation, and simplification. Its relevance lies in implementing RAG supported by the Unified Medical Language System (UMLS), a standardised medical terminology repository [51], which increases the reliability of generated responses. Although Ascle is not designed specifically for conversational agents or maternal health, it demonstrates a functional architecture in which the model retrieves relevant medical data before producing a response. Validation metrics such as ROUGE and BLEU are used to assess similarity between model outputs and reference answers, common benchmarks in biomedical QA evaluation [31]. In contrast, human expert validation that is focused on readability and relevance showed positive outcomes, while performance in accuracy and completeness was less satisfactory, indicating that the current model still requires refinement for high-stakes, complex applications.

The integration of RAG with LLMs has proven to be instrumental in mitigating issues like hallucination and knowledge staleness in biomedical environments. Notably, the MIRAGE benchmark and the MEDRAG tool enable systematic comparison across retrieval–generation configurations in medicine, achieving performance gains of up to 18% for models such as GPT-3.5 and Mixtral, narrowing the gap with GPT-4, by leveraging domain-specific corpora like PubMed and retrievers such as BM25 and MedCPT [52].

An evolution of traditional RAG is the Knowledge Graph-RAG (KG-RAG) approach, a low-cost, high-accuracy strategy that integrates language models with biomedical knowledge graphs such as SPOKE. This method improves context selection through minimal graph schemas and embedding techniques, reducing token consumption by more than 50% without sacrificing accuracy. KG-RAG significantly enhances the factual consistency and traceability of outputs by generating text that is grounded in explicit and verifiable knowledge structures [53].

Another significant advancement is BiomedRAG, which addresses noise limitations in standard RAG systems that retrieve content at the sentence or paragraph level. It introduces semantically relevant "document chunks" that improve precision in tasks like relation extraction and biomedical QA. Experiments across eight datasets demonstrated an average performance increase of 9.95%, with improvements over baselines reaching up to 4.97% [54].

i-MedRAG is a recent framework showcasing iterative reasoning for complex clinical questions. It enables LLMs to issue successive queries that iteratively retrieve evidence and refine responses, thereby building reasoning chains. This method achieved 69.68% accuracy in MedQA, outperforming prior fine-tuning and prompt-engineering techniques based on GPT-3.5 [55].

Two recent approaches also demonstrate improvements in explainability and precision using knowledge graphs. KRAGEN introduces advanced prompting strategies such as Graph-of-Thoughts (GoT) to decompose biomedical queries into subproblems, resolving them using contextualised knowledge-graph evidence. This reduces hallucinations and enhances explainability [56]. In parallel, MedGraphRAG incorporates triple structures and hierarchical retrieval (U-Retrieval), ensuring traceable, source-verified responses, which is especially effective for long-context medical tasks [57].

RAG systems and their advanced variants represent a major evolution in biomedical QA by combining language generation with retrieval mechanisms for structured, updatable knowledge. Techniques such as AMG-RAG and local graphs with multilevel summaries have shown improvements in precision and contextualisation in demanding benchmarks like MEDQA and MEDMCQA [58,59]. Recent studies also demonstrate that optimising retrieval depth and quality can strike a balance between performance, latency, and explainability, while graph-based retrievers enhance semantic coverage by capturing complex relationships that traditional embeddings miss [60]. Altogether, these findings reinforce RAG systems as a solid and scalable foundation for developing more accurate, reliable, and domain-adapted clinical tools.

Analysis of RAG systems and their advanced variants shows that their integration into biomedical QA addresses the need to combine generative capabilities with dynamic access to up-to-date knowledge. Unlike traditional retrieval methods such as BM25 or Dense Passage Retrieval (DPR), RAG systems have demonstrated superior accuracy and traceability in clinical contexts [3,33]. Moreover, recent studies highlight that specialized biomedical models do not always outperform general-purpose models on unseen clinical data, reinforcing the advantage of RAG as a bridge between generation and evidence-based retrieval [34].

The recent literature also indicates that fragmenting maternal clinical records as if they were generic biomedical prose can lead to inappropriate mixing of preconception notes, pregnancy observations, and postpartum records. This practice dilutes the urgency of certain findings, such as hypertensive symptoms after 20 weeks, when interpreted against pregestational baseline values [61]. To address this limitation, recent studies emphasize the need for episode-based information retrieval strategies, linked to unique identifiers for pregnancy, episode type, and gestational window, and complemented by the indexing of obstetric ontologies that restrict problems to the current pregnancy [62]. Furthermore, it is recommended that evaluation be stratified by trimester and postpartum period in order to reflect performance where it is most clinically relevant [63]. Collectively, these recommendations illustrate an emerging consensus on the importance of episode-based traceability and the integration of temporal constraints in RAG systems applied to maternal health records, ensuring that the information retrieved remains clinically consistent, contextually accurate, and directly useful for medical practice.

3.2. Axis 2: Development of LLMs Trained in Biomedical Domains

The rise of large language models (LLMs) has profoundly transformed natural language processing in the biomedical field, enabling complex tasks such as clinical report generation, relation extraction, and medical question answering to be automated with unprecedented efficiency. However, general-domain models often produce hallucinated or outdated responses, which poses particular risks in high-stakes domains like medicine [64]. To mitigate these limitations, domain-specific training of LLMs on the biomedical literature has gained traction. One example is BioGPT, which has shown significant improvements in specialised knowledge generation and extraction tasks [65]. In parallel, hybrid architectures like BiomedRAG integrate evidence retrieval with controlled generation, increasing the traceability and factual accuracy of outputs in clinical settings [54]. Additionally, models such as MedBioLM, which is fine-tuned on specific biomedical datasets, have shown meaningful improvements in both short- and long-form clinical reasoning tasks [66,67]. Together, these approaches form the foundation of a new generation of biomedical LLMs that arefocused on producing safer, more up-to-date, and clinically useful responses.

The recent literature has explored a variety of strategies to adapt LLMs to the needs of biomedical applications, aiming to improve precision, trustworthiness, and clinical relevance. A central approach has been the integration of information-retrieval mechanisms. For example, in [52], the MIRAGE benchmark was proposed to evaluate retrieval-augmented generation (RAG) systems in medicine. Using the MEDRAG toolkit, the authors demonstrated performance gains of up to 18% in models like GPT-3.5 by combining diverse biomedical corpora with specialised retrievers such as MedCPT and BM25, positioning PubMed as a key source for clinical QA.

Likewise, in [9], the authors introduced RAG2, a reasoning-guided architecture in which the model generates an intermediate justification used as a query for information retrieval. This strategy improves filtering and mitigates corpus bias, yielding performance increases of more than 5% over standard RAG across three closed-domain medical benchmarks [66]. MedBioLM, in turn, combines fine-tuning on biomedical datasets with RAG, improving factual coherence and performance in multiple-choice clinical QA tasks of both short and long formats.

Another notable contribution is PMC-LLaMA, an open-source biomedical model based on LLaMA. It is trained on 4.8 million research articles and 30,000 medical texts. Thanks to its instruction alignment and specialised corpus, PMC-LLaMA meets and even surpasses ChatGPT's performance in several medical benchmarks [68].

Complementarily, the Taiyi model demonstrates the value of bilingual fine-tuning using 140 datasets in both Chinese and English, covering more than ten biomedical task types. It stands out for its performance in multilingual and multitask biomedical QA [69]. From a critical perspective, Ref. [70] review QA approaches over electronic health records (EHRs), identifying strong dependence on emrQA, the limited availability of annotated datasets, and linguistic and regulatory barriers to real-world deployment. Similarly, Ref. [71] show that although ChatGPT-like models can perform biomedical classification and reasoning tasks, their performance remains inferior to domain-specialised models like BioBERT, particularly given the high computational cost of prompt engineering and the time required to reach optimal outcomes.

The development of biomedical language models that are trained on specialized corpora has enabled significant advances in the understanding of clinical terminology. However, evidence indicates that specialization alone does not guarantee superior performance across all contexts, highlighting the importance of fine-tuning and domain adaptation strategies [39]. In the case of maternal health, such adaptations include the integration of obstetric glossaries, international guidelines, and resources such as the OxMat dataset which reinforce the representativeness of diverse populations [7,72]. These techniques have been shown to improve the sensitivity of models for detecting critical warning signs, including preeclampsia and postpartum hemorrhage.

In general, advances in biomedical LLMs have produced promising tools for automating clinical and research tasks, yet their effective integration still faces several critical hurdles. Recent reviews emphasise that although zero-shot models can perform well in low-data scenarios, their performance is inconsistent and not always superior, which limits their general applicability [73]. Moreover, persistent challenges remain regarding explainability, data security, algorithmic bias, and hallucination risks (all of which compromise their use in sensitive clinical contexts [74,75]). Therefore, the adoption of specialised medical LLMs requires an interdisciplinary approach that combines evidence-based retrieval, rigorous evaluation in real-world scenarios, and ethical frameworks to ensure their trustworthiness and acceptance in clinical practice [74,76].

3.3. Axis 3: The Use of QA as Input for Clinical Conversational Agents

The integration of language models into QA systems has significantly transformed how medical information is accessed, processed, and communicated. In clinical contexts, conversational agents equipped with QA capabilities offer a promising avenue for enhancing patient education, supporting clinical decision-making, and reducing administrative burden [77]. Recent studies have demonstrated the potential of models like BioGPT and Med-PaLM 2 to generate medically accurate and comprehensible responses [78]. Moreover, initiatives based on human-centered design show that ethical and technical integration is key to the sustainable adoption of these agents [79]. Nonetheless, challenges remain related to answer explainability, adaptation to local contexts, and clinical validation.

Various recent works have explored the use of QA systems as functional backends for conversational agents in clinical environments. For instance, Ref. [80] presented a QA system based on BERT and GPT-2 that is trained on 5000 medical question–answer pairs. Although the system showed weaker quantitative performance (e.g., PPL, a standard metric for evaluating generative language models like GPT or BERT-GPT), it outperformed the base model qualitatively in terms of user intent comprehension.

Similarly, Ref. [81] introduced MedBot, a chatbot that uses natural language processing to deliver preliminary diagnoses and health recommendations based on user-input symptoms. The system enables users to describe their symptoms in natural language, then identifies possible conditions and suggests initial actions, such as lifestyle adjustments, symptom

monitoring, or consultation with a specialist. This makes MedBot particularly valuable in contexts with limited medical resources or where clinical access is costly or inefficient.

In another example, Ref. [82] proposed a biomedical assistant integrating the BioMistral model and RAG techniques. The system includes two key components: a semantic retrieval module that selects the most relevant document fragments, and a generative model (e.g., Llama2 or Mistral) that formulates coherent, context-aware responses. The primary objective is to enable users to pose complex clinical questions and receive relevant answers based on both general biomedical knowledge and personalised clinical documentation.

Ref. [83] conducted a study evaluating real patient interaction with a GPT-4-based chatbot. More than 50% of participants rated the chatbot's answers as better than those from search engines, although human physicians were still rated higher in comprehension and clarity. Interestingly, age did not significantly influence user preference or acceptance, suggesting the broad intergenerational accessibility of LLMs in clinical settings. Moreover, users with limited prior tech experience did not report major barriers in understanding or using the tool.

Ref. [84] developed an automated clinical information-extraction system deployed in a network of Italian hospitals. The system, named NEMT, uses a QA bot that is trained on thousands of real medical documents to convert unstructured clinical records into structured data. Leveraging advanced NLP and QA models, it achieved strong results, with an F1-score of 84.7% and an exact match accuracy of 78.1%, outperforming general-purpose models like ChatGPT-3.5. This approach not only automates administrative tasks but also enables large-scale data structuring for clinical research and decision support.

Another notable example is [85], who introduced MIA, a digital medical assistant for history-taking in radiology. It integrates a QA module, achieving 87% retrieval accuracy and an F1-score of 0.64. Evaluated in Swiss hospitals with real patients, the system was praised for its clarity and ease of use. However, areas for improvement were noted in conversational naturalness, data protection, and system security. MIA underscores the promise of QA-based agents in clinical care, provided that ethical considerations and user experience are carefully addressed.

The use of biomedical QA as input for clinical conversational agents represents a critical area for telemedicine. In obstetric care, where immediacy and clarity of information are essential, RAG-based systems have demonstrated the ability to integrate into clinical workflows without introducing significant delays or misinterpretations [72]. Recent reviews of telemedicine applications in obstetrics and gynecology further emphasize that these agents can enhance doctor–patient communication, expand access to care in rural settings, and reduce barriers to the availability of specialists [86].

QA systems have become foundational components in the development of clinical conversational agents, especially when integrated with advanced retrieval architectures and automated evaluation. The emergence of clinical benchmarks such as K-QA has enabled better measurement of answer completeness and reduced hallucinations in LLM outputs that are trained on real patient data, promoting their trustworthy use in healthcare settings [87]. Evaluation tools like ASTRID provide specific metrics to assess conversational fidelity, contextual relevance, and response accuracy in RAG-based QA systems, aligning closely with human clinical judgment [88]. Additionally, recent systematic reviews highlight that while QA-equipped conversational agents have shown potential in supporting healthcare professionals at the point of care, significant limitations remain, particularly regarding confidence and transparency evaluation and the lack of user-centered metrics [89]. Collectively, these advancements indicate that the use of QA as input for clinical agents is promising, especially when combined with high-quality datasets, rigorous evaluation, and robust ethical frameworks.

3.4. Axis 4: Methods for Clinical Validation and Explainability in QA Systems

In clinical contexts, QA systems powered by language models have gained popularity due to their potential to support medical decision-making. However, their deployment in real-world settings demands not only high accuracy but also rigorous clinical validation and reliable explainability. According to [90], explainability is essential for clinicians to trust artificial intelligence, particularly in high-risk tasks where errors may have severe consequences. Similarly, ref. [22] emphasise that the misalignment between model reasoning and human cognitive processes limits clinical acceptance. Recent studies such as [91] propose combined frameworks that integrate automatic validation with expert evaluation to address these challenges. Additionally, ref. [92] suggest that generating interpretable rationales alongside each answer can significantly improve clinical utility, encouraging ethical and transparent AI integration in healthcare.

Several approaches have been proposed for clinical validation and explainability in QA systems. A notable contribution by [93] presents a comprehensive framework based on six key dimensions of trust: factuality, robustness, fairness, safety, explainability, and calibration. This model not only defines individual criteria but also proposes an interconnected evaluation strategy that mirrors the complexity of clinical environments. Validation involves a combination of automated assessments and expert reviews to detect critical issues such as hallucinations or biases. Explainability is achieved through response traceability, grounding answers in medical sources, and generating rationales. Furthermore, evaluation findings inform model improvements, including adversarial fine-tuning and fact-verification modules. This framework offers a solid foundation for developing safer, more trustworthy clinical QA systems.

In parallel, ref. [94] introduced a hybrid architecture called GraphRAG, which combines symbolic reasoning with semantic retrieval, enhancing citation fidelity and transparency through knowledge graphs built with Neo4j and resources like UMLS. The system justifies answers by providing graph paths and explicit text fragments, offering not only accurate responses but also clear provenance for each claim. Unlike traditional QA systems that function as black boxes, this architecture incorporates query expansion and adaptive re-ranking to facilitate context-aware selection of clinical evidence. By enabling structured inspection of which medical entities and relationships support an answer, GraphRAG improves explainability and supports verifiable reasoning. While its overall accuracy is comparable to traditional models, its strength lies in documentation quality and auditability, which are critical in settings requiring transparency and accountability for medical automation.

Another study using the MedPAIR dataset explores the alignment between LLMs and human clinicians, revealing that models tend to prioritise different information than experts. The findings show that filtering irrelevant content based on human annotations improves performance for both models and humans [95]. MedPAIR introduces a novel evaluation framework that is focused on sentence-level contextual relevance, establishing metrics to assess alignment between human judgments and LLM-generated rationales. Techniques like ContextCite and self-reported relevance are used to quantify semantic agreement. These results are significant in clinical scenarios where interpretation depends on nuanced language or specific details in patient notes. The study suggests that explainability should go beyond showing evidence origin and ensure that the model processes information in ways that are aligned with expert reasoning. Thus, MedPAIR contributes to redefining QA validation criteria by integrating cognitive and clinical dimensions into AI evaluation.

NoteChat is a synthetic medical–patient dialogue system that is conditioned on clinical notes, using cooperative agents to improve response fluency, factuality, and reasoning. Human evaluations show over 20% improvement compared to models like ChatGPT and

GPT-4, highlighting its potential as a training base for explainable systems [96]. NoteChat's modular design, comprising planning, role-play, and refinement stages, produces more natural and clinically coherent dialogues, maintaining high alignment with structured medical content. Role-based interaction between LLM agents facilitates controlled reasoning and the generation of explicit rationales. This reduces hallucination risk and provides a scalable framework for evaluating and training QA models in clinical applications. From an explainability perspective, NoteChat enables traceable conversations in which each utterance can be audited and compared against the original medical note, offering a valuable tool for developing and validating QA systems with real-world medical utility.

Ref. [97] introduced a zero-shot prompting model from OpenAI, where answers are directly highlighted within clinical notes to promote traceability. This method facilitates seamless integration into clinical workflows, making answers not only useful but also auditable. Unlike free-text systems, this model visually marks relevant source segments, allowing for rapid verification by clinicians and fostering trust. The model includes visualisation mechanisms as direct evidence, aligning with transparency and verifiability standards. The study also simulates diverse clinical profiles to examine subjective perceptions of system usefulness, an essential factor in real-world adoption. Overall, the architecture not only delivers accurate responses but also ensures human interpretability, advancing QA systems that meet cognitive and operational demands in clinical practice.

In [98], the authors propose eKMQA, a medical reading comprehension framework that is capable of generating answers and corresponding rationales. It integrates reference texts with knowledge graphs and applies multitask learning to simultaneously train answer prediction and explanation generation, improving both accuracy and transparency. Unlike traditional models that prioritise correctness, eKMQA focuses on clinically valid rationales, offering explicit insights into model reasoning. Medical knowledge graphs support semantic relations among symptoms, diagnoses, and treatments, while natural language explanations help close the trust gap between AI and healthcare professionals. The framework also includes tools to evaluate explanation quality, measuring not only answer correctness but also the comprehensibility and usefulness of justifications for end users. In clinical environments where accountability and trust are essential, eKMQA provides a foundation for integrating AI systems ethically and safely into medical practice.

Overall, integrating QA systems into clinical settings requires a vision that combines accuracy, traceability, and alignment with medical reasoning. As the recent literature has demonstrated, accuracy alone is insufficient for safe adoption. For instance, ref. [99] highlights the need for adaptive trust-control mechanisms, especially in contexts where erroneous responses can have critical consequences. Likewise, ref. [100] argues that explainability must go beyond static visualisations; it should be contextual, interactive, and comprehensible to clinicians, fitting into their cognitive workflows. Ref. [101] stresses the importance of participatory evaluations with healthcare professionals, not just to assess technical performance but also clinical utility. In the same vein, ref. [102] suggests that algorithmic transparency and auditability are essential to uphold ethical principles in medical AI. Together, these perspectives reinforce that clinical validation and explainability are not final stages, but foundational pillars that should guide the design, development, and deployment of reliable, safe, and ethically sound clinical QA systems.

Clinical validation of QA systems requires not only technical performance testing but also systematic comparisons with human performance and safety assessments. Recent studies show that although biomedical models achieve high levels of accuracy, human professionals still outperform them in complex clinical tasks, reinforcing the need for robust, multimodal evaluation metrics [40]. The recent literature highlights a shift toward metrics that capture both factual accuracy and clinical safety, including precision, sensitivity in

detecting obstetric warning signs, and the critical error rate. These approaches address the need to evaluate not only the technical performance of models but also their potential impact on clinical practice, following validation frameworks proposed in recent studies on maternal health [39].

Although large language models (LLMs) have achieved strong results in medical question-answering tasks, human experts continue to surpass them in highly complex clinical scenarios [40]. However, recent evidence suggests that optimized retrieval-augmented generation (RAG) systems can significantly narrow this gap. For instance, the MIRAGE benchmark (2024) demonstrates that implementations such as MedRAG improve the accuracy of both general and specialized LLMs in medical QA tasks by up to 18%, reaching levels comparable to GPT-4 and, in some cases, outperforming pre-trained biomedical models on previously unseen datasets [52,103].

These findings support the competitiveness of RAG systems in tasks such as triage and the prediction of obstetric complications, underscoring their potential for safe and reliable integration into clinical practice. Building on this foundation, future evaluations should incorporate additional metrics that reflect not only factual accuracy but also the clinical safety and usability of the generated recommendations.

3.5. Axis 5: Specific Applications in the Domain of Maternal Health

In the field of maternal health, the application of biomedical QA faces the additional challenge of biases embedded in clinical data. Research conducted in Africa has shown that AI systems can replicate structural inequalities in the quality of care [104]. Similarly, it has been reported that biomedical algorithms exhibit gender biases that affect equity in access to health services [105]. Studies have documented that the evaluation of bias and fairness in biomedical models enables measurable improvements over baseline versions. Such findings support the recommendation to implement systematic audits in maternal health as a prerequisite for ethical and reliable deployment [5].

The recent literature on RAG systems in clinical contexts emphasizes that the safety of the assistance generated depends not only on the intrinsic quality of the model but also on the frequency of index updates and the end-to-end latency of queries [106,107]. In obstetric emergencies—such as preeclampsia, eclampsia, postpartum hemorrhage, or reduced fetal movement—delays in retrieving critical information can directly compromise patient safety. To mitigate these risks, recent studies recommend establishing explicit service objectives that include (i) an adequate indexing cadence to ensure that information remains current, (ii) maximum query latency thresholds under realistic system loads, and (iii) fallback strategies for situations in which index updates cannot be guaranteed [106]. Furthermore, it is suggested that the evaluation of medical alert systems incorporate specific safety-oriented metrics such as “alert time” and “failed escalation,” which go beyond overall accuracy and directly reflect clinical relevance [107]. These approaches enable the development of visualizations and indicators—such as those illustrated in Figure 3—that report on “update adherence” and “escalation sensitivity,” thereby aligning with best-practice recommendations for the safe deployment of RAG in maternal health.

Maternal health has undergone a profound transformation in recent years, driven by the increasing integration of digital technologies that enhance clinical care, improve data management, and promote equity in access to services. This progress is evident in various areas, from decision support systems and electronic health record (EHR) integration to predictive model optimisation and the adoption of innovative technologies that address sector-specific challenges.

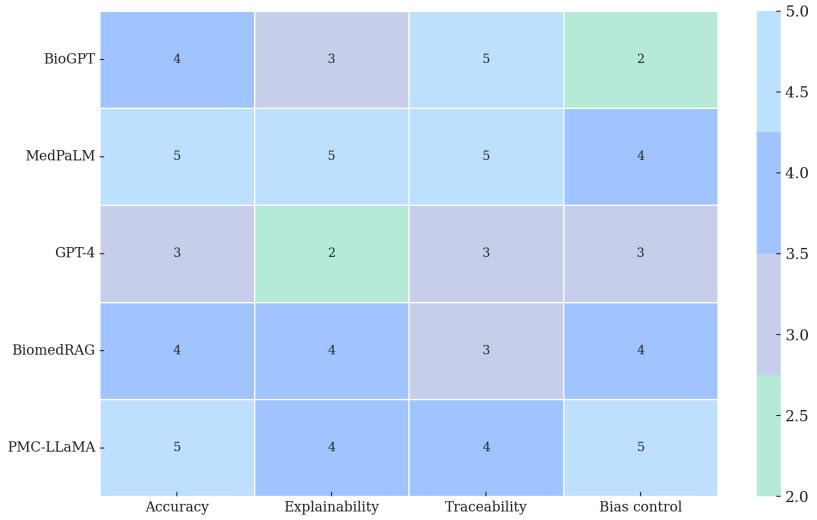

Figure 3. Heatmap comparing clinical validation attributes across different biomedical language models.

Maternal health is a cornerstone of medicine, involving a complex interplay of biological, social, and environmental factors affecting both the mother and the infant. Care during pregnancy, childbirth, and the postpartum period requires clinical decisions that are accurate, timely, and tailored to each patient. Technological advances have reshaped how such decisions are made, improving the quality and safety of care. However, this evolution also introduces challenges related to precision, integration, and accessibility that must be addressed to ensure continuous improvement in maternal health services [1].

EHRs have evolved from static data repositories into dynamic information sources for clinical decision support systems. Due to the volume and complexity of stored data, natural language processing (NLP) and text-mining methods can extract valuable insights from unstructured EHR fields such as clinical notes, obstetric histories, and examination reports. These techniques support the early identification of high-risk conditions like preeclampsia, preterm birth, and gestational diabetes, and they enable the generation of timely clinical alerts for remote care settings. When integrated with EHRs using retrieval-augmented generation (RAG), models such as MedPaLM, BioGPT, and GPT-4 have proven to be effective at extracting, summarising, and recommending actions based on clinical data [1,108].

Artificial intelligence has also enabled the development of targeted solutions within maternal health subdomains. For instance, in preeclampsia detection, machine learning and deep learning algorithms have outperformed conventional methods in sensitivity and specificity. Optimised models using particle swarm optimisation significantly reduce false negatives [109], a critical factor in managing this high-risk condition. Additionally, generative diffusion models have been used to synthesise ultrasound images for under-represented populations, improving diagnostic equity by training robust models even in data-scarce contexts such as rural or minority settings [110]. In fetal monitoring, recurrent neural networks process cardiotocographic signals to detect anomalies, enhancing response capabilities during acute events [111].

Semantic analysis of Spanish language clinical notes has been effective in detecting severe maternal morbidity. Locally trained NLP systems can automatically recognise clinical entities such as hypertension, proteinuria, bleeding, or neurological symptoms, identifying key obstetric events even in incomplete or informally written records. This capacity is especially valuable in high-demand or resource-limited regions, where AI can serve as a first layer of clinical surveillance [112].

The integration of AI with maternal EHRs represents a significant step forward in the digital transformation of obstetric care. This synergy enhances data collection, analysis, and utilization, enabling informed and personalized decisions for both clinicians and patients. Maternal EHRs have evolved beyond digital repositories to become dynamic tools, powered by AI, that can detect risk patterns, anticipate complications, and improve maternal and perinatal outcomes [113,114].

One of the main benefits of this integration is real-time data analysis, allowing intelligent systems to identify risk factors for severe maternal morbidity and other complications. At Fundación Hospital San Pedro in Colombia, an AI–EHR system monitors clinical data of pregnant patients, generating automatic alerts for potential risks and enabling timely preventive interventions and better resource allocation. The use of IoT devices further extends this monitoring to remote or underserved areas [114].

Interoperability is another key aspect of integrating AI and maternal EHRs. The adoption of international standards like HL7 FHIR enables seamless data sharing across institutions and systems, fostering continuity of care and collaborative research [115]. The U.S. National Institute of Child Health and Human Development has published FHIR-based guidelines to standardise maternal and child data exchange, supporting access to high-quality, longitudinal datasets for public health and clinical research [116]. These efforts address long-standing data fragmentation, ensuring accessibility and consistency across care networks.

Nevertheless, integrating AI and EHRs in maternal health poses challenges. Protecting the privacy and security of sensitive personal data is a top concern, particularly given the confidentiality of obstetric records [115,117]. Regulatory frameworks must safeguard data confidentiality, availability, consistency, and ethical use. Additionally, the transition to interoperable, AI-driven systems requires ongoing staff training and organisational adaptation to ensure effective, sustainable adoption [117]. Change management, data quality control, and clinical validation of predictive models are crucial in preventing errors and biases that could compromise patient safety and equity.

Optimising AI models in maternal health is essential to enhance diagnostic precision, treatment personalisation, and clinical management efficiency. This optimisation requires not only the development of advanced algorithms but also their adaptation to maternal populations, diverse data sources, and existing clinical systems. Effective AI integration is key to preventing and managing obstetric complications such as preeclampsia, preterm birth, and severe maternal morbidity [118].

Among the most effective strategies are supervised learning techniques combined with synthetic data augmentation. For example, generative adversarial networks (GANs) have been used to expand underrepresented datasets, improving model generalisability and reducing biases, an essential consideration for maternal health equity [119]. Model ensembles combining algorithms such as XGBoost and convolutional neural networks have reached AUC values above 0.9 for adverse event prediction [66,120].

Ref. [121] highlights federated learning as another key innovation. This approach trains AI models on distributed datasets across multiple hospitals without centralising information, preserving patient privacy while promoting robust, context-aware models. In Colombia, the MIA Colsubsidio platform incorporates real-time predictive analytics to anticipate maternal complications and optimise resource use, demonstrating the operational and clinical benefits of model optimisation [122].

Given the challenges of generative AI in telemedicine and prenatal care, more efficient models such as Small Language Models (SLMs) and fine-tuning techniques like QLoRA are being developed [123]. In highly localised applications, such as adapting a medical chatbot to specific populations, languages, or clinical practices, QLoRA offers a viable solution.

Unlike conventional fine-tuning methods that require vast GPU memory, QLoRA fine-tunes a 65B LLaMA model with under 48 GB of memory while maintaining comparable accuracy. A 7B model trained with QLoRA can run on just 5 GB, enabling deployment on smartphones or home health devices, which is a valuable asset for preventive and home-based telemedicine.

Practically, QLoRA enables small healthcare providers, universities, or startups to build or customise medical LLMs without enterprise-grade infrastructure. Its low memory footprint and estimated 72% carbon reduction in personal deployments make it both ethically and environmentally sustainable [123]. This innovation is transformative for generative AI in telemedicine, allowing the localised training of conversational clinical assistants that are adapted to specific needs, while ensuring privacy, reducing costs, and improving access.

The accuracy of AI models in maternal health is a key factor in their clinical utility and impact on maternal and perinatal outcomes. However, these models face technical and contextual challenges that limit generalisability. One major barrier is the variability and heterogeneity of clinical data across institutions, regions, and populations. Inconsistent documentation, especially unstructured or colloquial language in nursing notes, complicates automated processing and accurate information extraction [124].

Persistent inequalities in territorial, social, and gender dimensions affect maternal healthcare access and quality. According to the Pan American Health Organization, despite progress in reducing maternal mortality, rates remain high in several regions, particularly rural areas and among Indigenous and vulnerable populations [125,126]. These disparities stem from gaps in essential service coverage and workforce shortages. The International Council of Nurses reports that understaffing, especially of nurses and midwives, along with precarious working conditions, undermines care quality [127]. Generative AI can help mitigate these gaps by supporting clinical decision-making, continuous staff training, and remote patient engagement.

Additionally, generative AI can streamline maternal health administration and logistics, automating appointment scheduling, reminders, and personalised health education. This frees up professionals to focus on direct care and improves patient experience. However, strong digital infrastructure and staff training are needed for effective and sustainable implementation [128].

Lastly, the ethical and sustainable use of generative AI in maternal health demands a clear regulatory framework that protects patient rights, ensures algorithmic transparency, and promotes equitable access. Collaboration among governments, international agencies, academia, and private actors is essential to build inclusive, locally adapted policies, especially in settings facing humanitarian crises, climate change, or structural inequities [125].

4. Synthesis and Visual Analytics

To complement the narrative review and provide a multidimensional perspective on biomedical language models and their architectures, this section introduces three visual analyses derived from the literature. These representations synthesize trends in adoption, attributes related to clinical validation, and the comparative capabilities of different models. Their objective is to present the most relevant findings in a structured and interpretable format, thereby facilitating the identification of research patterns and highlighting critical areas for clinical practice [33,34,39].

4.1. Technology Prominence in the Literature

Figure 4 presents a horizontal bar chart synthesizing the frequency of citations of prominent biomedical language models and retrieval architectures within the reviewed

corpus. To construct this visualization, we systematically extracted citation counts from the final set of included references (2020–2025) and categorized them by model or architecture (e.g., BiomedRAG, BioGPT, MedPaLM, PMC-LLaMA). The counts were aggregated through a structured content analysis rather than automated bibliometrics, ensuring that only mentions directly linked to biomedical QA or retrieval-augmented applications in clinical or maternal health contexts were included.

BiomedRAG, BioGPT, and MedPaLM emerge as the most recurrently cited, reflecting their prominence as reference technologies in current biomedical QA research. The comparatively high citation frequency is not intended as a measure of clinical performance, but rather as an indicator of research interest, maturity, and perceived relevance across studies. This figure should therefore be interpreted as an interpretive bibliometric synthesis that highlights dominant research directions within the field.

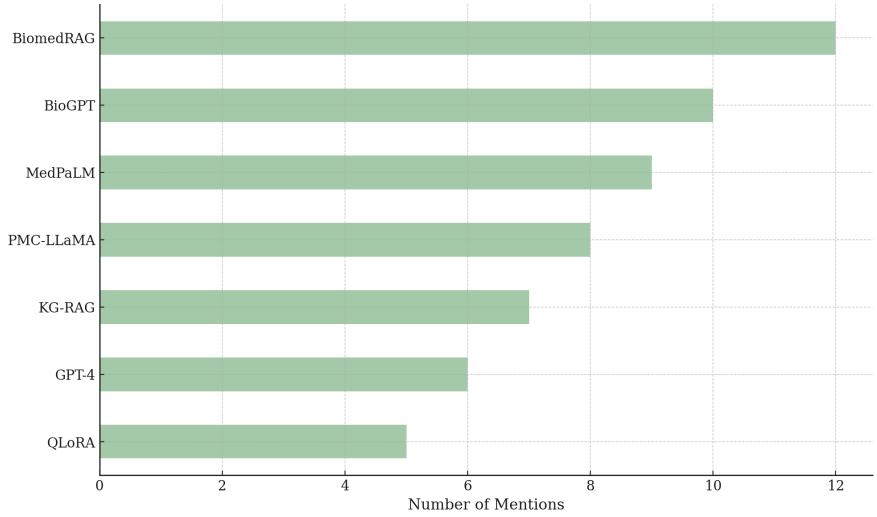

Figure 4. Frequency of citations for selected biomedical language models and retrieval architectures in the reviewed literature.

4.2. Clinical Validation Attributes by Model

To assess the clinical readiness of the reviewed technologies, Figure 3 presents a heatmap that synthesizes validation attributes for five biomedical language models across four dimensions: factual accuracy, explainability, traceability, and bias mitigation. The scores were derived through the structured content analysis of the included studies (2020–2025), where each model was evaluated based on qualitative or quantitative evidence reported in the literature. When multiple sources differed, scores were normalized by consensus weighting, prioritizing peer-reviewed benchmarks and domain-specific evaluations.

The visualization highlights that MedPaLM and PMC-LLaMA consistently exhibit the strongest validation profiles, particularly in explainability and traceability, two attributes that are essential for deployment in sensitive healthcare domains. GPT-4 demonstrates competitive performance in factual accuracy, yet its comparatively lower explainability raises concerns for applications requiring transparent decision support. Importantly, these values should be interpreted as an integrative synthesis of published evaluations, not as direct experimental measurements from this review.

Particularly in obstetrics, specialized clinical validation frameworks have been proposed to assess the performance of RAG models in scenarios such as obstetric triage and maternal complication prediction, underscoring the importance of context-adapted metrics [39].

4.3. Comparative Capabilities of Biomedical Language Models

Figure 5 provides a comparative overview of key capabilities synthesized from the recent literature. MedPaLM demonstrates balanced performance across all evaluated dimensions, while BiomedRAG stands out in traceability and explainability due to its retrieval-based architecture [3,34]. PMC-LLaMA also shows consistent performance, suggesting its suitability for integration into clinical pipelines with specific knowledge constraints. These comparisons are particularly relevant in telemedicine applications for maternal health, where the availability of reliable and explainable information is critical [52].

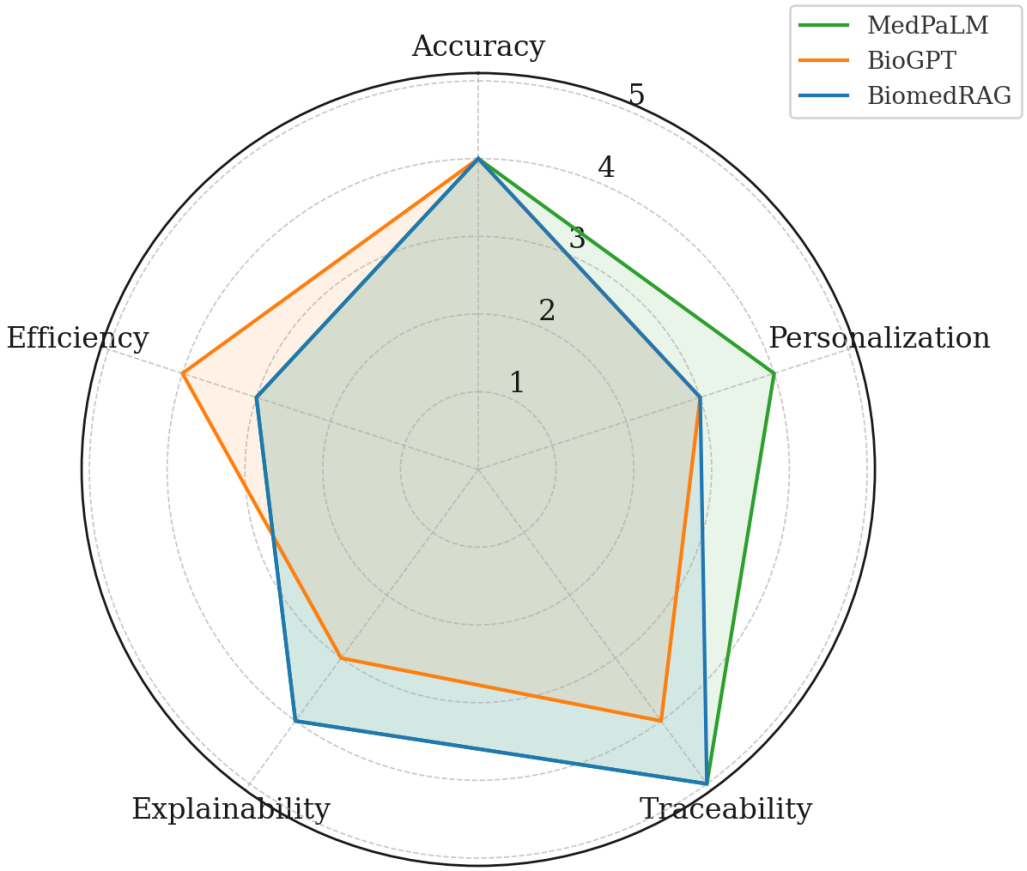

Figure 5. Radar chart comparing the capabilities of four biomedical language models across five key attributes: accuracy, personalization, traceability, explainability, and efficiency.

Taken together, these visualizations do not present original experimental results but rather an interpretive synthesis of the literature. They indicate that models with retrieval-based architectures tend to offer clear advantages in traceability and clinical safety, whereas generalist models excel in accuracy but face limitations in explainability. This contrast underscores the need for future research aimed at balancing technical performance with clinical requirements for effective adoption in maternal health and telemedicine [52].

These trends provide the foundation for the following discussion, which analyzes the challenges and opportunities associated with integrating these systems into real-world clinical settings, with a particular focus on maternal health.

5. Cross-Sectional Analysis and Discussion

The cross-sectional analysis of the five thematic axes reveals a strong convergence between the technical progress of generative AI systems, particularly QA systems, and the emerging clinical needs in maternal health. The integration of LLMs that are adapted to the

biomedical domain has been key in improving clinical language understanding, the personalisation of responses, and the traceability of generated recommendations. However, the effectiveness of these technologies depends on the careful integration of their components and the implementation of robust validation and explainability mechanisms.

First, RAG architectures and their variants (KG-RAG, BiomedRAG, GraphRAG) offer an effective response to the knowledge-staleness problem in purely generative models, mitigating risks such as hallucinations and enhancing contextual accuracy by enabling real-time access to the clinical literature [54,129]. These technical advances not only improve performance on benchmarks but also provide an operational framework that is more aligned with clinical practice, offering traceable, auditable, and up-to-date answers.

Advances in specialised LLMs such as PMC-LLaMA [68] and Taiyi [69] reflect a clear trend toward model specialisation by clinical domain, cultural context, and language. When combined with RAG pipelines, these models enable the development of more robust systems, as demonstrated in tasks such as information extraction, clinical summarization, and multi-step reasoning, which are critical capabilities in complex domains like maternal health.

The thematic axis that is focused on clinical conversational agents illustrates the role of QA as a functional module in telemedicine environments, contributing to improved accessibility, reduced administrative burden, and patient empowerment [81,84]. However, the literature also highlights the need for contextual validation, emphasising user-centered design, linguistic sensitivity, and participatory evaluation as critical factors for successful adoption and scalability.

Concerning clinical validation and explainability, there is a broad consensus on the need for frameworks that combine quantitative metrics (e.g., accuracy, F1-score) with expert clinical judgment and interpretable visualisation mechanisms. Initiatives such as MedPAIR [96], eKMQA [99], and NoteChat [97] advance toward more trustworthy AI by providing explicit rationales, alignment with medical reasoning, and semantic traceability. These proposals make clear that trust is not built on technical accuracy alone, but also on systems that align with the cognitive workflows of healthcare professionals.

The application of these technologies to maternal health consolidates it as a priority use case, where unequal access, the need for remote monitoring, and early risk detection converge to create an ideal context for deploying QA-assisted intelligent agents. The use of models that are trained on Spanish language clinical notes, the generation of synthetic data via GANs, and interoperability with standards like HL7-FHIR pave new paths for equity and continuous improvement in maternal care [112,115].

A recent systematic review on QA systems for healthcare professionals at the point of care [89] identified key barriers limiting clinical adoption: many solutions rely solely on unrealistic benchmarks, exhibit high bias risk, and lack mechanisms for communicating answer confidence or provenance. The authors recommend developing datasets that are representative of real-world clinical queries and implementing mechanisms to communicate not only answer accuracy but also confidence and traceability. This recommendation is especially relevant to QA systems in maternal health, where clinical decisions require transparency, traceability, and robustness, particularly in high-risk perinatal settings.

In the field of maternal–fetal medicine, AI has improved the accuracy and efficiency of diagnostic procedures such as perinatal ultrasound and fetal heart-rate monitoring [130]. According to a comprehensive 2023 review, AI-based models have reduced exploration time and clinician workload while improving the consistency of results. However, the authors caution that challenges remain concerning model interpretability and real-world clinical validation, underscoring the importance of involving clinicians at all stages of

development and evaluation. This aligns with the need for QA systems in maternal health that are explainable, auditable, and accepted by clinical staff.

6. Conclusions

This review synthesises the recent evolution of retrieval-augmented generation systems and domain-adapted large language models within the maternal-health landscape. Across the five thematic axes—RAG foundations, biomedical LLM advances, QA-driven conversational agents, clinical validation frameworks, and obstetric telehealth applications—the evidence converges on the feasibility of deploying traceable, context-aware question–answering pipelines that can operate safely in high-stakes perinatal settings. Architectures such as BiomedRAG, KG-RAG, and MedGraphRAG demonstrate measurable gains in factual accuracy and citation fidelity when compared with purely generative baselines. Meanwhile, specialized models like PMC-LLaMA, Med-PaLM 2, and Taiyi illustrate that domain pre-training and instruction tuning markedly improve clinical reasoning and multilingual coverage. Collectively, these developments position generative AI as a viable instrument for reducing informational asymmetries, supporting remote monitoring, and enhancing shared decision-making throughout pregnancy and the postpartum period.

Despite these advances, persistent challenges remain. The literature reveals unresolved risks related to biased training corpora, hallucination, and the variable alignment between model rationales and expert clinical reasoning. Furthermore, the performance of current systems continues to depend heavily on the quality of retrieval indices, the granularity of chunking strategies, and the timeliness of knowledge-base updates. Ethical integration in maternal–fetal medicine also demands transparent audit trails, robust governance over patient data, and safeguards to prevent the amplification of structural inequities, particularly in low-resource contexts where algorithmic decisions may disproportionately affect vulnerable populations.

To translate experimental success into sustainable clinical impact, future work should prioritise three lines of inquiry. First, federated evaluation networks need to be established so that models can be assessed on geographically diverse, privacy-preserving datasets while ensuring the reproducibility of results. Second, participatory co-design with obstetricians, midwives, and expectant mothers must inform interface design, dataset curation, and success metrics, thereby aligning technological affordances with user expectations and cultural norms. Third, energy- and memory-efficient optimisation techniques, exemplified by QLoRA and other parameter-efficient fine-tuning paradigms, should be adopted to facilitate on-device inference and equitable access in bandwidth-constrained settings.

The reviewed evidence confirms that retrieval-augmented biomedical LLMs hold substantial promise for maternal health when coupled with rigorous validation, transparent explainability, and context-sensitive deployment strategies. By advancing interdisciplinary collaborations and embedding ethical safeguards from inception to implementation, future research can foster intelligent maternal-care systems that are not only accurate and up-to-date but also trustworthy, inclusive, and adaptable to the realities of diverse clinical environments.

Author Contributions: Conceptualization, A.L.M.-B., M.D.N.-M., D.S.-V. and J.C.T.; methodology, A.L.M.-B., M.D.N.-M., J.C.T., S.R. and D.S.-V.; validation, A.N., S.R., D.S.-V. and J.C.T.; formal analysis, A.L.M.-B., M.D.N.-M. and A.N.; investigation, A.L.M.-B., S.R. and M.D.N.-M.; resources, A.N.; data curation, A.L.M.-B. and M.D.N.-M.; writing—original draft preparation, A.L.M.-B., M.D.N.-M., A.N., S.R., D.S.-V. and J.C.T.; writing—review and editing, A.L.M.-B., M.D.N.-M., A.N., S.R., D.S.-V. and J.C.T.; visualization, A.L.M.-B. and M.D.N.-M.; supervision, S.R., A.N.; project administration, S.R., A.N.; funding acquisition, S.R., A.N. All authors have read and agreed to the published version of the manuscript.

Funding: This article is derived from the project "Intelligent agent based on natural language processing for maternal monitoring in the post-pandemic era within a telemedicine environment," Code 82244, aimed at strengthening research capacities and developing solutions in the field of artificial intelligence applied to telemedicine contexts. This initiative is part of the project developed under the framework of Minciencias Call 890 of 2020, and it currently receives financial support from the Ministry of Science, Technology and Innovation of Colombia—Minciencias, under Contract RC No. 2023-0678. This support has been essential for the formulation, development, and dissemination of the results presented in this manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding authors.

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

Use of Artificial Intelligence: During the preparation of this work, the authors used ChatGPT 4o and Grammarly v1.2 to improve the writing. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

1. 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. [CrossRef] [PubMed]
2. Gaber, F.; Shaik, M.; Allega, F.; Bilecz, A.J.; Busch, F.; Goon, K.; Franke, V.; Akalin, A. Evaluating large language model workflows in clinical decision support incorporating RAG on real-world cases. *npj Digit. Med.* **2025**, *8*, 16. [CrossRef] [PubMed]
3. Amugongo, L.M.; Mascheroni, P.; Brooks, S.; Doering, S.; Seidel, J. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLoS Digit. Health* **2025**, *4*, e0000877. [CrossRef] [PubMed]
4. Ozmen, B.B.; Mathur, P. Evidence-based artificial intelligence: Implementing retrieval-augmented generation models to enhance clinical decision support in plastic surgery. *J. Plast. Reconstr. Aesthetic Surg.* **2025**, *104*, 414–416. [CrossRef]
5. Ji, Y.; Zhang, H.; Wang, Y. Evaluating bias in retrieval-augmented medical question-answering systems. *arXiv* **2025**, arXiv:2503.15454. [CrossRef]
6. Lin, M.; Lin, L.; Lin, L.; Lin, Z.; Yan, X. A bibliometric analysis of the advance of artificial intelligence in medicine. *Front. Med.* **2025**, *12*, 1504428. [CrossRef]
7. Khan, M.J.; Duta, I.; Albert, B.; Cooke, W.; Vatish, M.; Jones, G.D. The OxMat dataset: A multimodal resource for the development of AI-driven technologies in maternal and newborn child health. *arXiv* **2024**, arXiv:2404.08024. [CrossRef]
8. Park, C.; Moon, H.; Park, C.; Lim, H. MIRAGE: A Metric-Intensive Benchmark for Retrieval-Augmented Generation Evaluation. *arXiv* **2025**, arXiv:2504.17137. [CrossRef]
9. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S. On the opportunities and risks of foundation models. *Nat. Mach. Intell.* **2022**, *4*, 189–191. [CrossRef]
10. World Health Organization. *Pulse Survey on Continuity of Essential Health Services During the COVID-19 Pandemic*; WHO: Geneva, Switzerland, 2020. Available online: https://www.who.int/publications/i/item/WHO-2019-nCoV-EHS_continuity-survey-2020.1 (accessed on 7 February 2025).
11. Ohannessian, E.; Duong, A.; Odone, A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: A call to action. *JMIR Public Health Surveill.* **2020**, *6*, e18810. [CrossRef]
12. Portnoy, M.A. Telemedicine in the COVID-19 era: A balancing act to avoid harm. *J. Allergy Clin. Immunol. Pract.* **2020**, *8*, 2459–2461. [CrossRef]
13. Dávila, L.S.; Rivera, R.R.; Tapia, J.H.; Asanza, W.R. Inteligencia artificial aplicada a la oftalmología: ResNet-50 y VGG-19 en el diagnóstico de catarata y glaucoma. *Inform. Sist. Rev. Tecnol. Inform. Las Comun.* **2024**, *8*, 52–59. [CrossRef]
14. Schwab, K. *The Fourth Industrial Revolution*; World Economic Forum: Geneva, Switzerland, 2016. Available online: https://www.weforum.org/about/the-fourth-industrial-revolution-by-klaus-schwab (accessed on 8 February 2025).
15. Topol, E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*; Basic Books: New York, NY, USA, 2019.
16. World Health Organization. *WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening*; WHO: Geneva, Switzerland, 2019. Available online: https://www.who.int/publications/i/item/9789241550505 (accessed on 8 February 2025).

17. Agarwal, S.; LeFevre, A.E.; Lee, J.; L'Engle, K.; Mehl, G.; Sinha, C.; Labrique, A. Guidelines for reporting of health interventions using mobile phones: Mobile health (mHealth) evidence reporting and assessment (mERA) checklist. *BMJ* **2016**, *352*, i1174. [CrossRef] [PubMed]
18. Ramakrishnan, R.; Rao, S.; He, J.-R. Perinatal health predictors using artificial intelligence: A review. *Women's Health* **2021**, *17*, 17455065211046132. [CrossRef]
19. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223. [CrossRef] [PubMed]
20. Bhattarai, K.; Oh, I.Y.; Sierra, J.M.; Tang, J.; Payne, P.R.O.; Abrams, Z.; Lai, A.M. Leveraging GPT-4 for identifying cancer phenotypes in electronic health records: A performance comparison between GPT-4, GPT-3.5-turbo, Flan-T5, Llama-3-8B, and spaCy's rule-based and machine learning-based methods. *JAMIA Open* **2024**, *7*, ooae060. [CrossRef]
21. Mahyoub, M.; Dougherty, K.; Shukla, A. Extracting pulmonary embolism diagnoses from radiology impressions using GPT-4o: Large language model evaluation study. *JMIR Med. Inform.* **2025**, *13*, e67706. [CrossRef]
22. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
23. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [CrossRef]
24. Ktena, I.; Wiles, O.; Albuquerque, I.; Rebuffi, S.-A.; Tanno, R.; Roy, A.G.; Azizi, S.; Belgrave, D.; Kohli, P.; Cemgil, T.; et al. Los modelos generativos mejoran la equidad de los clasificadores médicos en los cambios de distribución. *Nat. Med.* **2024**, *30*, 1166–1173. [CrossRef]
25. Bedi, S.; Liu, Y.; Orr-Ewing, L.; Dash, D.; Koyejo, S.; Callahan, A.; Fries, J.A.; Wornow, M.; Swaminathan, A.; Lehmann, L.S.; et al. A Systematic Review of Testing and Evaluation of Healthcare Applications of Large Language Models (LLMs). *medRxiv* **2024**. [CrossRef]
26. Gargari, O.K.; Habibi, G. Mejora de la IA médica con generación aumentada por recuperación: Una mini revisión narrativa. *Digit. Health* **2025**, *11*, 1–7. [CrossRef]
27. Zheng, Y.; Yan, Y.; Chen, S.; Cai, Y.; Ren, K.; Liu, Y.; Zhuang, J.; Zhao, M. Integración de la generación aumentada de recuperación para mejorar las recomendaciones personalizadas de los médicos en los servicios médicos basados en la web: Estudio de desarrollo de modelos. *Front. Public Health* **2025**, *13*, 1501408. [CrossRef]
28. Bora, A.; Cuayáhuitl, H. Systematic analysis of retrieval-augmented generation-based LLMs for medical chatbot applications. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 2355–2374. [CrossRef]
29. Li, Y.; Shen, X.; Yang, C.; Cao, Z.; Du, R.; Yu, M.; Wang, J.; Wang, M. Novel electronic health records applied for prediction of pre-eclampsia: Machine-learning algorithms. *Pregnancy Hypertens. Int. J. Women's Cardiovasc. Health* **2021**, *26*, 102–109. [CrossRef] [PubMed]
30. Ge, J.; Sun, S.; Owens, J.; Galvez, V.; Gologorskaya, O.; Lai, J.C.; Pletcher, M.J.; Lai, K. Development of a Liver Disease-Specific Large Language Model Chat Interface using Retrieval Augmented Generation. *medRxiv* **2023**. [CrossRef]
31. Chen, X.; Zhang, W.; Zhao, Z.; Xu, P.; Zheng, Y.; Shi, D.; He, M. ICGA-GPT: Report generation and question answering for indocyanine green angiography images. *Br. J. Ophthalmol.* **2024**, *1208*, 1450–1456. [CrossRef]
32. K2View. What Are AI Hallucinations? Available online: https://www.k2view.com/what-are-ai-hallucinations/ (accessed on 2 June 2025).
33. Kumar, A.; Lee, S.; Park, J. Adoption of biomedical large language models: A scoping review of applications and challenges. *J. Biomed. Inform.* **2024**, *157*, 104703. [CrossRef]
34. Dorfner, F.J.; Dada, A.; Busch, F.; Makowski, M.R.; Han, T.; Truhn, D.; Kleesiek, J.; Sushil, M.; Lammert, J.; Adams, L.C.; et al. Biomedical large language models seem not to be superior to generalist models on unseen medical data. *arXiv* **2024**, arXiv:2408.13833. [CrossRef]
35. Carvallo, M.; Peso, J.; Zapata-Toloza, R.; Andalaft, C. Telehealth and telemedicine in Latin America: A scoping review. *Salud Cienc. Tecnol.* **2025**, *5*, 1185. [CrossRef]
36. Dirección General de Comunicación Social, UNAM. Mayor Relevancia de la Telemedicina en Atención Durante el Embarazo (Boletín UNAM-DGCS-712). Universidad Nacional Autónoma de México. Available online: https://www.dgcs.unam.mx/boletin/bdboletin/2021_712.html (accessed on 29 August 2021).
37. Valencia, S.A.; Barrientos, J.G.; Silva, E.A.T.; Díaz, E.S. Impacto en los resultados en salud de la telesalud aplicada para la atención y seguimiento ambulatorio del alto riesgo obstétrico: Revisión narrativa de la literatura. *Med. UPB* **2024**, *43*, 43–51. [CrossRef]
38. NUBIX. (s. f.). La Teleginecología y sus Beneficios en la era de la Telemedicina. Available online: https://nubix.cloud/radiologia/la-teleginecologia-y-sus-beneficios-en-la-era-de-la-telemedicina (accessed on 1 February 2025).
39. Gargari, O.K.; Habibi, G. Enhancing medical AI with retrieval-augmented generation: A mini narrative review. *Digit. Health* **2025**, *11*, 20552076251337177. [CrossRef]

40. Wan, N.; Jin, Q.; Chan, J.; Xiong, G.; Applebaum, S.; Gilson, A.; McMurry, R.; Taylor, R.A.; Zhang, A.; Chen, Q.; et al. Humans Continue to Outperform Large Language Models in Complex Clinical Decision-Making: A Study with Medical Calculators. *arXiv* **2024**, arXiv:2411.05897. [CrossRef]
41. Yang, R.; Zeng, Q.; You, K.; Qiao, Y.; Huang, L.; Hsieh, C.-C.; Rosand, B.; Goldwasser, J.; Dave, A.; Keenan, T.; et al. Ascle—Un kit de herramientas de procesamiento del lenguaje natural de Python para la generación de textos médicos: Estudio de desarrollo y evaluación. *J. Med. Internet Res.* **2024**, 26, e60601. [CrossRef]
42. Tecnoloblog. Qué es un Sistema RAG y cómo Funciona: Guía Exhaustiva y Actualizada. Available online: https://www.tecnoloblog.com/sistemas-rag/ (accessed on 8 February 2025).
43. León, M.C.C.; Núñez, y.J.E.R. Análisis de Modelos de Inteligencia Artificial Aplicados a Sistemas Biomédicos e Internet de Objetos Médicos. Universidad Politécnica Salesiana. 2024. Available online: https://dspace.ups.edu.ec/bitstream/123456789/27874/1/UPS-GT005362.pdf (accessed on 5 February 2025).
44. Aliste, F.A. INTELIGENCIA ARTIFICIAL GENERATIVA: LLMS en Medicina. SECOIR. 2025. Available online: https://secoir.org/wp-content/uploads/2025/05/10.7-Monografia-SECOIR-2025-V1.pdf (accessed on 5 March 2025).
45. Guanoluisa, J.M.; Chicaiza, R.P.M.; Avalos, C.J.B. Agente Conversacional para Consultas Sobre Servicio Médico en una Clínica Privada. *3C Tecnol.* **2021**, *10*, 47–71. Available online: https://dialnet.unirioja.es/descarga/articulo/8044473.pdf (accessed on 5 March 2025). [CrossRef]
46. Aloy-Duch, A.; Vila, M.S.; Ramos-D'Angelo, F.; Calo, L.A.; Llaneza-Velasco, M.E.; Fortuny-Organs, B.; Apezetxea-Celaya, A. Desarrollo y Validación de Estándares para Unidades de Calidad de Centros Sanitarios. *J. Healthc. Qual. Res.* **2023**, *38*, 366–375. Available online: https://www.elsevier.es/es-revista-journal-healthcare-quality-research-257-articulo-desarrollo-validacion-estandares-unidades-calidad-S260364792300057X (accessed on 15 February 2025). [CrossRef]
47. Capasso, A.; de Mucio, B.; Ramírez, D.; Colomar, M.; Serruya, Y.S. Salud Digital en Salud Materna: Avances y Desafíos en América Latina y el Caribe. OPS. 2024. Available online: https://www.paho.org/es/noticias/7-3-2024-salud-digital-salud-materna-avances-desafios-america-latina-caribe (accessed on 8 October 2025).
48. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
49. Zhang, Y. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv* **2023**, arXiv:2309.01219. [CrossRef]
50. Joshi, S. Retrieval Augmented Generation for Medical Question-Answering with Llama-2–7b. Medium. Available online: https://medium.com/@sauravjoshi23/retrieval-augmented-generation-for-medical-question-answering-with-llama-2-7b-82486847d089 (accessed on 6 March 2025).
51. Bodenreider, O. The Unified Medical Language System (UMLS). National Library of Medicine. Available online: https://www.nlm.nih.gov/research/umls/index.html (accessed on 15 July 2025).
52. Xiong, G.; Jin, Q.; Lu, Z.; Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. In *Findings of the Association for Computational Linguistics: ACL*; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 6233–6251. Available online: https://teddy-xionggz.github.io/benchmark-medical-rag/ (accessed on 5 March 2025).
53. Soman, K.; Rose, P.W.; Morris, J.H.; Akbas, R.E.; Smith, B.; Peetoom, B.; Villouta-Reyes, C.; Cerono, G.; Shi, Y.; Rizk-Jackson, A.; et al. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics* **2024**, *40*, btae560. [CrossRef]
54. Li, M.; Kilicoglu, H.; Xu, H.; Zhang, R. BiomedRAG: A retrieval augmented large language model for biomedicine. *J. Biomed. Inform.* **2025**, *162*, 104769. [CrossRef]
55. Xiong, G.; Jin, Q.; Wang, X.; Zhang, M.; Lu, Z.; Zhang, A. Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*; World Scientific: Singapore, 2025.
56. Matsumoto, N.; Moran, J.; Choi, H.; Hernandez, M.E.; Venkatesan, M.; Wang, P.; Moore, J.H. KRAGEN: A Knowledge Graph-Enhanced RAG Framework for Biomedical Problem Solving Using Large Language Models. *Bioinform. Adv.* **2024**, *40*, btae353. Available online: https://github.com/EpistasisLab/KRAGEN (accessed on 8 February 2025). [CrossRef]
57. Wu, J.; Zhu, J.; Qi, Y.; Chen, J.; Xu, M.; Menolascina, F.; Grau, V. Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation. 2024. Available online: https://github.com/MedicineToken/Medical-Graph-RAG (accessed on 8 February 2025).
58. Rezaei, M.R.; Fard, R.S.; Parker, J.; Krishnan, R.G.; Lankarany, M. Adaptive Knowledge Graphs Enhance Medical Question Answering: Bridging the Gap Between LLMs and Evolving Medical Knowledge. *arXiv* **2025**, arXiv:2502.13010. [CrossRef]
59. Guan, L.; Huang, Y.; Liu, J. Biomedical Question Answering via Multi-Level Summarization on a Local Knowledge Graph. *arXiv* **2025**, arXiv:2504.01309. [CrossRef]
60. Delile, J.; Mukherjee, S.; Van Pamel, A.; Zhukov, L. Graph-Based Retriever Captures the Long Tail of Biomedical Knowledge. *arXiv* **2024**, arXiv:2402.12352. [CrossRef]

61. Lyu, T.; Liang, C.; Liu, J.; Campbell, B.; Hung, P.; Shih, Y.; Ghumman, N.; Li, X.; Haendel, M.A.; Chute, C.G. Temporal Events Detector for Pregnancy Care (TED-PC): A rule-based algorithm to infer gestational age and delivery date from electronic health records of pregnant women with and without COVID-19. *PLoS ONE* **2022**, *17*, e0276923. [CrossRef]
62. Zhu, Y.; Ren, C.; Xie, S.; Liu, S.; Ji, H.; Wang, Z.; Sun, T.; He, L.; Li, Z.; Zhu, X. REALM: RAG-Driven Enhancement of Multimodal Electronic Health Records Analysis via Large Language Models. *arXiv* **2024**, arXiv:2402.07016. [CrossRef]
63. Zhao, Z.; Yuan, H.; Liu, J.; Chen, H.; Ying, H.; Zhou, S.; Yu, S. Evaluating Entity Retrieval in Electronic Health Records: A Semantic Gap Perspective. *arXiv* **2025**, arXiv:2502.06252. [CrossRef]
64. He, J.; Zhang, B.; Rouhizadeh, H.; Chen, Y.; Yang, R.; Lu, J.; Chen, X.; Liu, N.; Li, I.; Teodoro, D. Retrieval-Augmented Generation in Biomedicine: A Survey of Technologies, Datasets, and Clinical Applications. *arXiv* **2025**, arXiv:2505.01146. [CrossRef]
65. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.-Y. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *arXiv* **2022**, arXiv:2210.10341. [CrossRef]
66. Kim, S. MedBioLM: Optimizing Medical and Biological QA with Fine-Tuned Large Language Models and Retrieval-Augmented Generation. *arXiv* **2025**, arXiv:2502.03004. [CrossRef]
67. Sohn, J.; Park, Y.; Yoon, C.; Park, S.; Hwang, H.; Sung, M.; Kim, H.; Kang, J. Rationale-Guided Retrieval Augmented Generation for Medical Question Answering. 2024. Available online: https://github.com/dmis-lab/RAG2 (accessed on 8 February 2025).
68. Wu, C.; Lin, W.; Zhang, X.; Zhang, Y.; Xie, W.; Wang, Y. PMC-LLaMA: Toward Building Open-Source Language Models for Medicine. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1833–1843. [CrossRef]
69. Luo, L.; Ning, J.; Zhao, Y.; Wang, Z.; Ding, Z.; Chen, P.; Fu, W.; Han, Q.; Xu, G.; Qiu, Y.; et al. Taiyi: A bilingual fine-tuned large language model for diverse biomedical tasks. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1865–1874. [CrossRef]
70. Bardhan, J.; Roberts, K.; Wang, D.Z. Question Answering for Electronic Health Records: Scoping Review of Datasets and Models. *J. Med. Internet Res.* **2024**, *26*, 53636. [CrossRef]
71. Chen, S.; Li, Y.; Lu, S.; Van, H.; Aerts, H.J.; Savova, G.K.; Bitterman, D.S. Evaluating the ChatGPT family of models for biomedical reasoning and classification. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 940–948. [CrossRef] [PubMed]
72. Vani, M.S.; Sudhakar, R.V.; Mahendar, A.; Ledalla, S.; Radha, M.; Sunitha, M. Personalized health monitoring using explainable AI: Bridging trust in predictive healthcare. *Sci. Rep.* **2025**, *15*, 31892. [CrossRef]
73. Laskar, I.J.; Peng, C.; Huang, J. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Comput. Biol. Med.* **2024**, *171*, 108189. [CrossRef] [PubMed]
74. Wang, C.; Li, M.; He, J.; Wang, Z.; Darzi, E.; Chen, Z.; Ye, J.; Li, T.; Su, Y.; Ke, J.; et al. A survey for Large Language Models in Biomedicine. *arXiv* **2024**, arXiv:2409.00133. [CrossRef]
75. Ullah, E.; Parwani, A.; Baig, M.M.; Singh, R. Diagnostic Pathology Team. Challenges and barriers of using large language models such as ChatGPT for diagnostic medicine with a focus on digital pathology: A scoping review. *Diagn. Pathol.* **2024**, *19*, 1464. [CrossRef]
76. Yu, L.; Fan, L.; Li, L.; Zhou, J.; Ma, Z.; Xian, L.; Hua, W.; He, S.; Jin, M.; Zhang, Y.; et al. Large Language Models in Biomedical and Health Informatics: A Review with Bibliometric Analysis. *arXiv* **2024**, arXiv:2403.16303. [CrossRef]
77. Spatharou, A.; Hieronimus, S.; Jenkins, J. Transforming Healthcare with AI: The Impact on the Workforce and Organizations, McKinsey & Company. 2020. Available online: https://www.mckinsey.com/industries/healthcare/our-insights/transforming-healthcare-with-ai (accessed on 8 October 2025).
78. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Towards expert-level medical question answering with Med-PaLM 2. *Nature* **2023**, 620, 113–122. [CrossRef]
79. Li, D.; Williams, P.; Wang, W.; Sahay, S. Towards building ethical and safe conversational agents for health applications. *ACM Trans. Comput.-Hum. Interact. (TOCHI)* **2021**, *28*, 1–36. [CrossRef]
80. Jeong, S.W.; Kim, C.G.; Whangbo, T.K. Question Answering System for Healthcare Information based on BERT and GPT. In Proceedings of the 2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Phuket, Thailand, 22–25 March 2023; pp. 1–6.
81. Anjum, K.; Sameer, M.; Kumar, S. AI Enabled NLP based Text to Text Medical Chatbot. In Proceedings of the 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM), Uttar Pradesh, India, 22–24 February 2023; pp. 1–6.
82. Esther, C.; Kanisshka, U.P.; Ananya, G.S.; Tamizhmalar, D.; Elangovan, V.; Ishan Raghavender, N. Biomedical Chat Assistant with Personalized Document Reader Using BioMistral and RAG. In Proceedings of the 2025 International Conference on Computing and Communication Technologies (ICCCT), Chennai, India, 16–17 April 2025; pp. 1–6.
83. Carl, N.; Haggenmüller, S.; Wies, C.; Nguyen, L.; Winterstein, J.T.; Hetz, M.J.; Mangold, M.H.; Hartung, F.O.; Grüne, B.; Holland-Letz, T.; et al. Evaluating interactions of patients with large language models for medical information. *BJU Int.* **2025**, *135*, 1010–1017. [CrossRef] [PubMed]

84. Crema, C.; Verde, F.; Tiraboschi, P.; Marra, C.; Arighi, A.; Fostinelli, S.; Giuffré, G.M.; Dal Maschio, V.P.; L'abbate, F.; Solca, F.; et al. Medical Information Extraction With NLP-Powered QABots: A Real-World Scenario. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 6906–6918. [CrossRef]
85. Denecke, K.; Reichenpfader, D.; Willi, D.; Kennel, K.; Bonel, H.; Nairz, K.; Cihoric, N.; Papaux, D.; von Tengg-Kobligk, H. Person-based design and evaluation of MIA, a digital medical interview assistant for radiology. *Front. Artif. Intell.* **2024**, *7*, 1431156. [CrossRef]
86. Medani, I.E.; Hakami, A.M.; Chourasia, U.H.; Rahamtalla, B.; Adawi, N.M.; Fadailu, M.; Salih, A.; Abdelmola, A.; Hashim, K.N.; Dawelbait, A.M.; et al. Telemedicine in Obstetrics and Gynecology: A Scoping Review of Enhancing Access and Outcomes in Modern Healthcare. *Healthcare* **2024**, *13*, 2036. [CrossRef]
87. Manes, I.; Ronn, N.; Cohen, D.; Ber, R.I.; Horowitz-Kugler, Z.; Stanovsky, G. K-QA: A Real-World Medical Q&A Benchmark. *arXiv* **2024**, arXiv:2401.14493. [CrossRef]
88. Chowdhury, M.; He, Y.V.; Higham, A.; Lim, E. ASTRID–An Automated and Scalable TRIaD for the Evaluation of RAG-based Clinical Question Answering Systems. *arXiv* **2025**, arXiv:2501.08208. [CrossRef]
89. Kell, G.; Roberts, A.; Umansky, S.; Qian, L.; Ferrari, D.; Soboczenski, F.; Wallace, B. Question answering systems for health professionals at the point of care–A systematic review. *arXiv* **2024**, *31*, 1009–1024. [CrossRef]
90. Sadeghi, Z.; Alizadehsani, R.; CIFCI, M.A.; Kausar, S.; Rehman, R.; Mahanta, P.; Bora, P.K.; Almasri, A.; Alkhawaldeh, R.S.; Hussain, S.; et al. A review of Explainable Artificial Intelligence in healthcare. *Comput. Electr. Eng. Int. J.* **2024**, *118*, 109370. [CrossRef]
91. Chiatti, A.; Bernardini, S.; Piccolo, L.S.G.; Schiaffonati, V.; Matteucci, M. Mapping user trust in Vision Language Models: Research landscape, challenges, and prospects. *arXiv* **2025**, arXiv:2505.05318. [CrossRef]
92. Li, H.; Chen, Z.; Zhang, J.; Wang, Q. Rationale-enhanced clinical QA with multi-level explanation generation. *Artif. Intell. Med.* **2024**, *142*, 102570. [CrossRef]
93. Wang, Y.; Mercer, R.E.; Rudzicz, F.; Roy, S.S.; Ren, P.; Chen, Z.; Wang, X. Trustworthy medical question answering: An evaluation-centric survey. *arXiv* **2025**, arXiv:2506.03659. [CrossRef]
94. Sekar, T.; Kushal, K.; Shankar, S.; Mohammed, S.; Fiaidhi, J. Investigations on using evidence-based GraphRAG pipeline using LLM tailored for answering USMLE medical exam questions. *medRxiv* **2025**. [CrossRef]
95. Hao, Y.; Alhamoud, K.; Jeong, H.; Zhang, H.; Puri, I.; Torr, P.; Schaekermann, M.; Stern, A.D.; Ghassemi, M. MedPAIR: Measuring physicians and AI relevance alignment in medical question answering. *arXiv* **2025**, arXiv:2505.24040. [CrossRef]
96. Wang, J.; Yao, Z.; Yang, Z.; Zhou, H.; Li, R.; Wang, X.; Xu, Y.; Yu, H. NoteChat: A dataset of synthetic doctor-patient conversations conditioned on clinical notes. *arXiv* **2024**, arXiv:2310.15959.
97. Albassam, D. Toward human-centered interactive clinical question answering system. *arXiv* **2025**, arXiv:2505.18928.
98. Li, D.; He, S.; Hu, B.; Chen, Q. Towards explainable medical machine reading comprehension with rationale generation. *IEEE Trans. Audio Speech Lang. Process.* **2025**, *33*, 1675–1683. [CrossRef]
99. Mehrtash, A.; Wells, W.M., III; Tempany, C.M.; Abolmaesumi, P.; Kapur, T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 3868–3878. [CrossRef]
100. Müller, V.; Reichert, C.; Scheuermann, H. Explainability in Clinical AI: Designing Transparent Decision Support Tools. *Artif. Intell. Med.* **2021**, *117*, 102111. [CrossRef]
101. Ji, M.; Genchev, G.Z.; Huang, H.; Xu, T.; Lu, H.; Yu, G. Evaluation framework for successful artificial intelligence-enabled clinical decision support systems: Mixed methods study. *J. Med. Internet Res.* **2021**, *23*, e25929. [CrossRef]
102. Khosravi, M.; Zare, Z.; Mojtabaeian, S.M.; Izadi, R. Artificial intelligence and decision-making in healthcare: A thematic analysis of a systematic review of reviews. *Health Serv. Res. Manag. Epidemiol.* **2024**, *11*, 23333928241234863. [CrossRef]
103. Xiong, G. MedRAG: A Systematic Toolkit for Retrieval-Augmented Generation on Medical Question Answering. GitHub. 2024. Available online: https://github.com/Teddy-XiongGZ/MedRAG (accessed on 7 February 2025).
104. Ncube, M. Incomplete Chronicles: Unveiling Data Bias in Maternal Health. Mozilla Foundation. 2024. Available online: https://www.mozillafoundation.org/en/research/library/incomplete-chronicles-unveiling-data-bias-in-maternal-health/ (accessed on 10 March 2025).
105. Joshi, A. Big data and AI for gender equality in health: Bias is a big challenge. *Front. Big Data* **2024**, *7*, 1436019. [CrossRef] [PubMed]
106. Neha, F.; Bhati, D.; Shukla, D.K. Retrieval-Augmented Generation (RAG) in healthcare: A comprehensive review. *AI* **2025**, *6*, 226. [CrossRef]
107. Yuan, S.; Yang, Z.; Li, J.; Wu, C.; Liu, S. AI-Powered early warning systems for clinical deterioration significantly improve patient outcomes: A meta-analysis. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 203. [CrossRef]
108. Hu, T.; Zhou, X.-H. Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions. *arXiv* **2024**, arXiv:2404.09135. [CrossRef]

109. Boga, Z.; Sándor, C.; Kovács, P. A Multidimensional Particle Swarm Optimization-Based Algorithm for Brain MRI Tumor Segmentation. *Sensors* **2025**, *25*, 2800. [CrossRef]
110. Gao, Z.; Peromingo Peromingo, D.; Cubillo Romero, J. Generación de Imágenes Mediante Modelos de Difusión. 2024. Available online: https://docta.ucm.es/entities/publication/66fa4de2-e0b2-4d2b-88cf-3885523dc16e (accessed on 10 March 2025).
111. Park, S.H. Artificial intelligence for ultrasonography: Unique opportunities and challenges. *Ultrasonography* **2021**, *40*, 3–6. [CrossRef] [PubMed] [PubMed Central]
112. Yan, Y.; Wang, K.; Feng, B.; Yao, J.; Jiang, T.; Jin, Z.; Zheng, Y.; Zhou, Y.; Chen, C.; Sui, L.; et al. The use of large language models in detecting Chinese ultrasound report errors. *npj Digit. Med.* **2025**, *8*, 66. [CrossRef]
113. Centro de Investigación de la Universidad San Agustín (USAT). La Inteligencia Artificial como Desafío en la Salud Materna Perinatal. USA. 2023. Available online: https://www.usat.edu.pe/articulos/la-inteligencia-artificial-como-desafio-en-la-salud-materna-perinatal/ (accessed on 20 February 2025).
114. INS–Instituto Nacional de Salud. *Protocolo de Vigilancia en Salud Pública de Morbilidad Materna Extrema*; Versión 4; Instituto Nacional de Salud: Bogotá, Colombia, 2022. [CrossRef]
115. Consultorsalud. Interoperabilidad de la Historia Clínica Electrónica. 2021. Available online: https://consultorsalud.com/interoperabilidad-de-la-historia-clinica-electronica/ (accessed on 15 March 2025).
116. Eunice Kennedy Shriver National Institute of Child Health and Human Development. Guía Materna: Artículo de Interés: Una guía Desarrollada por el NICHD Establece un Marco para Vincular los Datos de Salud Materna y Salud Infantil. NIH. 2023. Available online: https://espanol.nichd.nih.gov/noticias/prensa/062623-guia-materna (accessed on 26 June 2023).
117. ¿Cuáles Son Los Retos De Implementar Inteligencia Artificial En Los Sistemas De Salud Y Cómo Manejarlos Eficientemente? Atlantis University. 2024. Available online: https://atlantisuniversity.edu/es/au_blog/retos-inteligencia-artificial-en-salud/ (accessed on 15 June 2025).
118. Atlantis University. La Inteligencia Artificial en el Sector Salud ¿Están en Riesgo Algunos Trabajos? AU Blog. 2023. Available online: https://atlantisuniversity.edu/es/au_blog/inteligencia-artificial-para-sector-salud/ (accessed on 1 March 2025).
119. Mhatre, A.; Warhade, S.R.; Pawar, O.; Kokate, S.; Jain, S.; Emmanuel, M. Leveraging LLM: Implementing an Advanced AI Chatbot for Healthcare. *Int. J. Innov. Sci. Res. Technol.* **2024**, *9*, 3144–3151. [CrossRef]
120. Torres, L.F. XGBoost: The King of Machine Learning Algorithms | by Luís Fernando Torres | LatinXinAI | Medium. Medium. 2023. Available online: https://medium.com/latinxinai/xgboost-the-king-of-machine-learning-algorithms-6b5c0d4acd87 (accessed on 8 February 2025).
121. Abbas, S.R.; Abbas, Z.; Zahir, A.; Lee, S.W. Federated Learning in Smart Healthcare: A Comprehensive Review on Privacy, Security, and Predictive Analytics with IoT Integration. *Healthcare* **2024**, *12*, 2587. [CrossRef]
122. Salud Colsubsidio Lanza el Primer modelo de Patología 100% Digital en Colombia y Transforma el Diagnóstico Clínico en el país. Available online: https://consultorsalud.com/salud-colsubsidio-modelo-patologia-digital/ (accessed on 15 June 2025).
123. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. NeurIPS. 2023. Available online: https://github.com/artidoro/qlora (accessed on 8 February 2025).
124. Mapari, S.A.; Shrivastava, D.; Dave, A.; Bedi, G.N.; Gupta, A.; Sachani, P.; Kasat, P.R.; Pradeep, U. Revolutionizing Maternal Health: The Role of Artificial Intelligence in Enhancing Care and Accessibility. *Cureus* **2024**, *16*, e69555. [CrossRef] [PubMed] [PubMed Central]
125. OPS Conmemora el Día Mundial de la Salud Destacando Avances y Desafíos en salud Materna y Neonatal en la Región-OPS/OMS | Organización Panamericana de la Salud. Available online: https://www.paho.org/es/noticias/7-4-2025-ops-conmemora-dia-mundial-salud-destacando-avances-desafios-salud-materna (accessed on 15 June 2025).
126. Organización Panamericana de la Salud (OPS). La OPS destaca avances en la reducción de la mortalidad materna en las Américas, pero advierte sobre desafíos persistentes. *OPS/OMS* **2025**, b56.
127. El CIE se une a la OMS para Priorizar la Salud Materna y del Recién Nacido Mediante la Inversión en Enfermería | ICN-International Council of Nurses. ICN-International Council of Nurses. Available online: https://www.icn.ch/es/noticias/el-cie-se-une-la-oms-para-priorizar-la-salud-materna-y-del-recien-nacido-mediante-la (accessed on 15 June 2025).
128. UNFPA América Latina y el Caribe | Salud Materna. UNFPA LAC. Available online: https://lac.unfpa.org/es/topics/salud-matern (accessed on 15 June 2025).
129. Chen, Y.; Zhang, W.; Liu, K. GraphRAG: Enhancing Biomedical QA with Knowledge Graph Grounding. *J. Biomed. Inform.* **2024**, *154*, 104217.
130. Qiao, S.; Fang, X.; Garrett, C.; Zhang, R.; Li, X.; Kang, Y. Generative AI for qualitative analysis in a maternal health study: Coding in-depth interviews using Large Language Models (LLMs). *medRxiv* **2024**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.