future internet

MDPI

Review

Trends, Challenges, and Applications of Large Language Models in Healthcare: A Bibliometric and Scoping Review

Vincenza Carchiolo *,† and Michele Malgeri *,†

Dipartimento Ingegneria Elettrica Elettronica e Informatica, Università di Catania, Via Santa Sofia 64, 95125 Catania, Italy

* Correspondence: vincenza.carchiolo@unict.it (V.C.); michele.malgeri@unict.it (M.M.)

† These authors contributed equally to this work.

Abstract: The application of Large Language Models (*LLM*s) in medicine represents an area of growing interest in scientific research. This study presents a quantitative review of the scientific literature aiming at analyzing emerging trends in the use of *LLM*s in the medical field. Through a systematic analysis of works extracted from Scopus, the study examines the temporal evolution, geographical distribution, and scientific collaborations between research institutions and nations. Furthermore, the main topics addressed in the most cited papers are identified, and the most recent and relevant reviews are explored in depth. The quantitative approach enables mapping the development of research, highlighting both opportunities and open challenges. This study presents a comprehensive analysis of research articles and review-type articles across several years, focusing on temporal, geographical, and thematic trends. The temporal analysis reveals significant shifts in research activity, including periods of increased or decreased publication output and the emergence of new areas of interest. Geographically, the results identify regions and countries with higher concentrations of publications, as well as regions experiencing growing or stagnant international collaboration. The thematic analysis highlights the key research areas addressed in the reviewed papers, tracking evolving topics and changes in research focus over time. Additionally, the collaborative analysis sheds light on key networks of international collaboration, revealing changes in the distribution of affiliations across subperiods and publication types. Finally, an investigation of the most cited papers highlights the works that have had the greatest impact on the scientific community, identifying enduring themes and methodologies that continue to shape the field of study. The results provide a clear overview of current trends and future perspectives for the application of *LLM*s in medicine, offering a valuable reference for researchers and professionals in the field.

Keywords: large language models; artificial intelligence; healthcare; e-health; deep learning; transformer; neural networks; review; bibliometric analysis

Academic Editor: Petros Patias

Received: 18 December 2024
Revised: 24 January 2025
Accepted: 5 February 2025
Published: 8 February 2025

Citation: Carchiolo, V.; Malgeri, M. Trends, Challenges, and Applications of Large Language Models in Healthcare: A Bibliometric and Scoping Review. *Future Internet* **2025**, *17*, 76. https://doi.org/10.3390/fi17020076

1. Introduction

The aim of this paper is to explore an emerging and complex topic: the use of large language models (*LLM*) in the field of e-health through the analysis of publications. The main goal is to provide a general overview, identifying the main areas of interest and the most relevant research trends, also adding some geographical and topological deepening.

An *LLM* is a type of artificial intelligence model, specifically a language model, trained on vast amounts of textual data. These models use deep learning techniques, particularly transformer neural networks, to analyze, understand, and generate natural language text. *LLM* are designed to predict words, sentences, or paragraphs and are capable of performing

multiple tasks, including answering questions on various topics, generating coherent and creative texts (such as articles, stories, or poems), translating different languages, and summarizing long documents. More advanced applications also extend to understanding and analyzing the sentiment of a text.

The use of large language models (*LLM*s) in healthcare is rapidly transforming various aspects of medical practice and research. Numerous benefits have been achieved using *LLM*s in healthcare, particularly with time savings by automating documentation and repetitive tasks. Additionally, a significant advantage is the ability to support diagnostics by integrating clinical guidelines and extensive datasets, leading to an enhancement of results. Finally, *LLM*s enable the delivery of health information and services remotely through AI-driven applications. Large language models (*LLM*s) are transforming healthcare by offering innovative solutions across a range of applications. One key area is Clinical Decision Support, where *LLM*s provide evidence-based recommendations, diagnostic suggestions, and treatment options, enhancing clinical efficiency and accuracy [1]. Another impactful use is in Medical Documentation and Summarization, where *LLM*s automate documentation tasks by transcribing physician–patient interactions and generating structured summaries, significantly reducing the administrative burden on healthcare providers [2]. *LLM*s also contribute to improved Patient Interaction and Virtual Assistants, powering chatbots and virtual health assistants that offer general health information, schedule appointments, and send medication reminders [3,4]. In the field of Personalized Medicine, these models analyze patient-specific data to deliver tailored treatment recommendations based on genetic, clinical, and lifestyle information [5]. Lastly, Health Education benefits from *LLM*s by creating interactive educational content that simplifies complex medical concepts for both professionals and patients [6]. Together, these applications illustrate the transformative potential of *LLM*s in healthcare, though challenges in privacy, bias, and interpretability remain critical for future advancements. However, several challenges remain, especially in the domains of data privacy and security, bias and fairness, and the interpretability of results, which remains a critical issue. A scoping review on *LLM*s in the healthcare field is essential to outline the potential and challenges of this emerging technology, supporting informed and responsible implementation in clinical practice and research. This approach allows for exploring the scope, nature, and reach of the existing literature, identifying the main areas of application (such as diagnosis, education, and patient support), and emerging trends in research. *LLM*s present both promises and risks. On the one hand, they can improve clinical efficiency, support decision making, and reduce errors. On the other hand, they raise ethical issues, data bias concerns, and patient safety worries. A scoping review helps to clarify these dynamics, providing a solid foundation for new research questions dealing with health, computer science, and social matters.

There are several reviews in the literature on the use of *LLM* in healthcare, but they often focus on specific topics. For example, ref. [7] focused on creating models specific to medicine, ref. [6] examined the use of *LLM*s in medical education, ref. [8] addressed applications limited to image analysis, ref. [9] explored ethical considerations, and and others have presented a general taxonomy, but focused on specific models like ChatGPT [10,11]. However, none of these reviews provide an overall view of the temporal development of the topic or a geographical analysis of how interest in the use of *LLM* in healthcare has evolved, as proposed in [12]. This paper aims to fill this gap by presenting an analysis of the temporal, geographical, and item trends in interest for *LLM*s in the healthcare field. In a second phase, the most relevant works will be analyzed to highlight the key themes of interest, from the perspective of technological solutions and practical applications. For these purposes, classical techniques from natural language processing, artificial intelligence,

and complex network analysis will be used. To this end, data extracted from Scopus [13] will be used, and several distinct datasets will be created to examine various aspects of the topic.

This study presents a comprehensive analysis of research articles and review-type articles across several years, focusing on temporal, geographical, and thematic trends. The temporal analysis reveals significant shifts in research activity, including periods of increased or decreased publication output and the emergence of new areas of interest. Geographically, the results identify regions and countries with higher concentrations of publications, as well as regions experiencing growing or stagnant international collaboration. The thematic analysis highlights the key research areas addressed in the reviewed papers, tracking evolving topics and changes in research focus over time. Additionally, the collaborative analysis sheds light on key networks of international collaboration, revealing changes in the distribution of affiliations across subperiods and publication types. Finally, an investigation of the most cited papers highlights the works that have had the greatest impact on the scientific community, identifying enduring themes and methodologies that continue to shape the field of study.

In Sections 2 and 3, we briefly introduce *LLM*s and the methodologies and technologies used for the review. Section 4 introduces the structures of the datasets while in Section 5, the results of the quantitative analysis are presented, and some of the most significant papers are discussed. Finally, in Section 6, some conclusions are drawn, and potential future improvements are outlined.

2. Overview of Large Language Models and Their Taxonomy

An *LLM* is a sophisticated artificial intelligence capable of understanding and producing text in a manner very similar to a human being. These models are trained on vast amounts of textual data, thereby absorbing grammatical rules, vocabulary, and the nuances of language. The operation of *LLM*s is based on a complex mechanism of machine learning. A key innovation at the heart of modern *LLM*s is the attention mechanism, which enables the model to dynamically focus on different parts of the input text, assigning varying levels of importance to each word or token based on context. This mechanism underpins transformer architectures, where multiple attention heads work in parallel to capture intricate dependencies between words, making these models highly effective for natural language tasks. By analyzing immense amounts of text, these models identify patterns and relationships between words, learning to predict which words best fit a given sentence or context. At the core of this capability lies the use of neural networks—computational systems inspired by the human brain—that enable *LLM*s to learn and make predictions. Once trained, *LLM*s can perform a wide range of tasks: from generating entirely new texts to automatic translation, from answering complex questions to writing code. Essentially, *LLM*s are versatile tools that open new frontiers in human–machine interaction. In summary, *LLM*s are AI models that, thanks to machine learning the attention mechanism and neural networks, can understand and generate text in increasingly sophisticated and natural ways.

The roots of *LLM*s trace back to the 1950s and 1960s, with the development of the first artificial neural networks. These networks, inspired by the functioning of the human brain, offered a new perspective for tackling complex problems such as natural language understanding. However, it was only with the advent of backpropagation in the 1980s that neural networks began to become practical tools for complex tasks. This algorithm allowed networks to learn from their mistakes and improve their performance over time. In recent years, the explosion of *LLM*s has been made possible by increasingly powerful hardware and the advent of cloud computing, which have enabled the training of larger

and more complex models. Another fundamental aspect contributing to the rise of *LLM*s is the availability of vast amounts of data. This availability is one of the most significant effects of the explosion of the internet and digitalization, which have become the fuel for training *LLM*s.

One of the main characteristics of an *LLM* is, indeed, the size of the model and the quantity of data used during the training phase. An *LLM* contains an enormous number of parameters, which are the weights and biases that the model learns during training. The greater the number of parameters, the higher the model's ability to capture complex information from the text. The largest models contain billions of parameters. *LLM*s are trained on massive datasets, which may include documents, newspaper articles, books, web pages, and more. This training allows them to learn a wide range of information about language, grammar, context, and facts, which are influenced by key characteristics of an *LLM* such as its architecture, the number of parameters (in billions), and the size of the training corpus.

Figure 1 shows the main *LLM*s based on their characteristics and the type of architecture. The temporal evolution highlights a significant increase in both the number of parameters and the size of the training datasets, particularly for the most recent models.

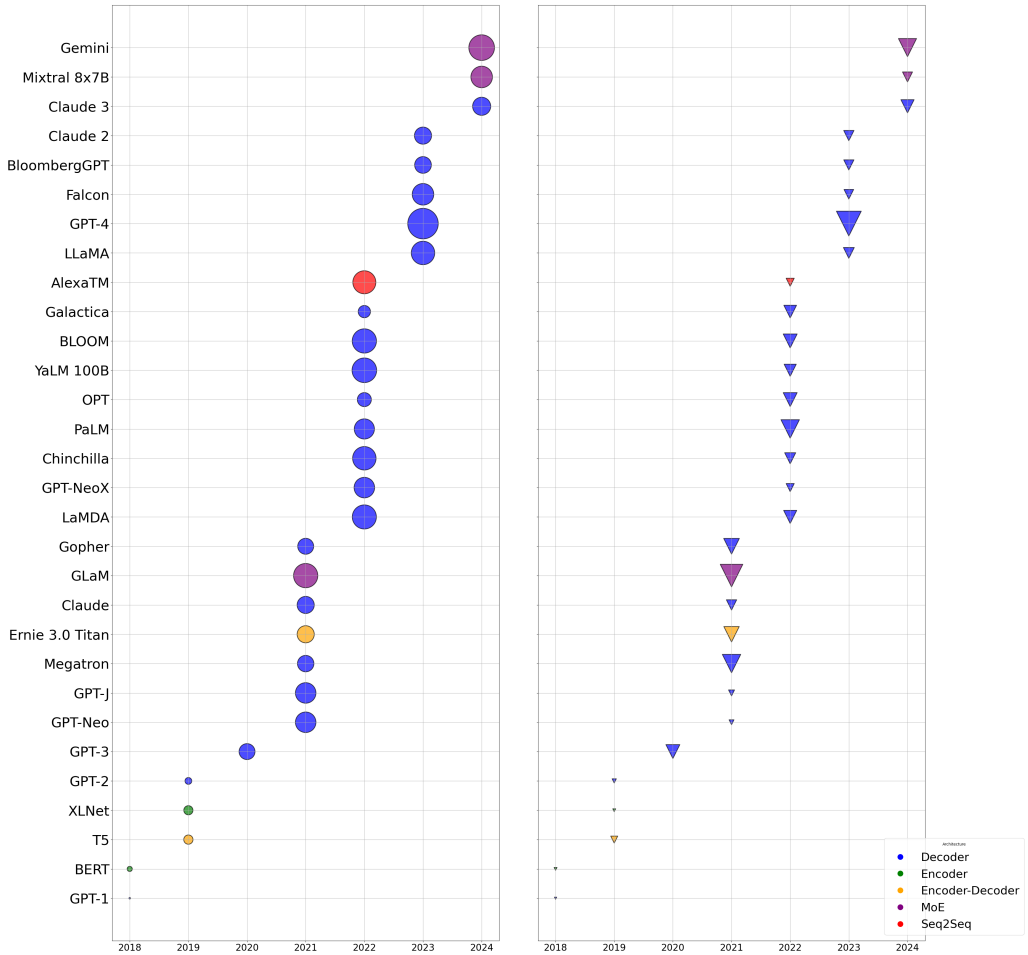


Figure 1. *LLM* by year. X-axis (Year): Indicates the introduction period of each model. Y-axis (model name, parameters in billions—B): Logarithmic scale representing the number of parameters in each model. Point size: Proportional to the size of the training dataset (in trillions of tokens). Triangle size: Proportional to the number of model parameters (in billions). Colors represent different *LLM* architectures.

The most recent models, such as GPT-4 [14], are trained on much larger volumes of data compared to their predecessors, reflecting the increasing availability and use of data to enhance model capabilities. The growth in the number of parameters is evident over time, with models like GPT-4 and GLaM [15] having orders of magnitude more parameters than earlier models such as GPT-1 or BERT [16].

One of the most significant features of *LLM*s is their ability to address various aspects of natural language processing. Below is a list highlighting the key capabilities they cover:

- *Text completion:* Generating sentences or paragraphs from a given input;
- *Translation:* Translating text from one language to another;
- *Question answering:* Answering questions based on context or prior knowledge;
- *Summarization:* Summarizing long texts;
- *Content creation:* Writing articles, stories, essays, or code;
- *Conversations:* Simulating conversations with users, like chatbots.

*LLM*s represent one of the most significant advancements in the field of AI, due to their ability to handle and understand natural language with levels of accuracy and fluency never seen before. A list of some of the most well-known models follows:

1. *GPT-3 [17]* and *GPT-4*, by OpenAI, are among the most famous language models. GPT-3, for example, has 175 billion parameters and is capable of generating highly complex and coherent texts.
2. *BERT* (Bidirectional Encoder Representations from Transformers) was designed by Google and is distinguished by its ability to consider the context both before and after a word, making it highly effective for tasks such as understanding meaning in sentences.
3. *T5* (Text-To-Text Transfer Transformer) [18] transforms all natural language processing (NLP) tasks into a text generation format, making it extremely versatile.
4. *PaLM* (Pathways Language Model) [19]: Another model developed by Google with billions of parameters, aimed at competing with GPT in various linguistic tasks.
5. *Mistral* [20] is another notable language model that has gained attention for its innovative approach. Developed by a team of researchers focused on advancing AI, Mistral is designed to improve the efficiency and flexibility of natural language understanding. Unlike traditional models, Mistral uses a unique architecture that optimizes training efficiency while maintaining high performance across a wide range of NLP tasks. Mistral features billions of parameters, similar to models like GPT-3 and PaLM, but with a focus on achieving optimal performance using a more compact model. This allows Mistral to perform complex tasks with relatively less computational resource usage, making it an attractive choice for real-time applications and systems with limited computational power. In addition to text completion, summarization, and translation, Mistral has shown impressive capabilities in tasks such as semantic analysis, context-aware content generation, and complex question answering. It also supports domain-specific applications, allowing customization for specialized fields like healthcare or legal sectors. By combining an advanced transformer architecture with a focus on computational efficiency, Mistral has quickly emerged as a strong competitor in the field of large language models, delivering results comparable to other well-established models while offering potential advantages in terms of performance scalability and energy efficiency.

*LLM*s can be classified based on different models used for training, with some of the most popular being as follows.

Reinforcement Learning from Human Feedback (RLHF) [21] uses human feedback to improve the responses of an *LLM*, aligning the model with desired behaviors. Learning

occurs through reinforcement algorithms such as Proximal Policy Optimization (PPO). This technique has been used in many cases; for example, OpenAI uses RLHF in advanced versions of ChatGPT to reduce inappropriate responses or align the model with ethical values. DeepMind's Sparrow [22] provides reliable and safe responses. This solution can be helpful if a model initially provides unethical or misleading responses, as human feedback can guide learning to correct them, such as responding to sensitive questions or preventing bias.

Instruction Tuning aims to make the model responsive to prompts with clear instructions, with the goal of improving the model's ability to follow specific instructions using a set of human-generated examples. This approach makes the model more reactive. For example, Google trained the T5 model with instruction tuning (FLAN-T5), allowing it to better respond to prompts with specific instructions.

Mixture of Experts (MoE) divides a model into subnetworks ("experts"), each specialized in different tasks. During inference, only some experts are activated, reducing computational load. Examples of this method include Google's Generalist Language Model (GlaM) [15], Switch Transformer, and Mistral. For example, GLaM uses MoE to activate only a part of the model during inference, reducing computational costs.

Despite the various architectures and training techniques used by different language models, one common issue across all of them is hallucinations. This phenomenon occurs when a model generates responses that appear plausible but are actually incorrect or entirely fabricated. Hallucinations arise because language models, whether autoregressive like GPT, encoder–decoder like T5, or models like BERT, rely on statistical correlations between data rather than true semantic understanding. While these models are capable of learning complex linguistic structures, their lack of deep content comprehension often leads to the generation of inaccurate or non-existent information, especially when dealing with rare data or data that were not adequately represented during training. This issue is a common challenge for all large-scale language models and represents one of the main areas of research aimed at improving their reliability and ability to generate correct and verified responses.

*LLM*s are finding promising applications in the medical field in various areas:

- *Diagnostic support*: *LLM*s can analyze clinical data and assist in the diagnostic process by suggesting possible diagnoses or treatment plans based on symptom descriptions;
- *Natural language interpretation*: They can help translate clinical notes or medical reports into structured information useful for analysis;
- *Medical question answering*: *LLM*s can answer complex questions related to medications, treatments, or clinical conditions;
- *Assistance in research*: They can help researchers find relevant information in large volumes of scientific articles or generate summaries of studies.

GPT, Mistral, and BERT are widely used in medical and general natural language processing applications. GPT has often been employed for medical chatbots, clinical decision support, and scientific text generation. It can summarize documents, answer medical questions, and assist in medical training through simulations. Mistral, with its focus on efficiency and high performance, is valuable for applications in medical text analysis and the automation of healthcare documentation processes. BERT is frequently used for text classification, information extraction, and natural language understanding in clinical data. These models, along with their specialized variants (see Table 1), play a crucial role in advancing artificial intelligence in medicine, enhancing diagnostic accuracy, clinical support, and access to medical knowledge.

Table 1. Main *LLM*s in the medical domain with purposes and key features

Model	Developer	Purpose	Features
BioGPT [23]	Microsoft Research	Biomedical NLP	Trained on biomedical literature for text generation and clinical question answering
GatorTron [24]	University of Florida	Clinical data analysis	Trained on millions of electronic medical notes to improve clinical language understanding
MedPaLM [25]	Google Health	Diagnosis and medical Q&A	Combines PaLM with medical data to answer clinical questions
PubMedBERT [26]	—	Biomedical NLP	BERT version optimized for PubMed articles, useful for text classification and information extraction
Galactica [27]	Meta AI	Scientific research	Trained on scientific texts, including medical publications
ClinicalBERT [28]	—	Clinical language understanding	Adaptation of BERT to electronic health records (EHR)
BioBERT [29]	—	Biology and medical applications	Trained on PubMed and PMC to enhance NLP accuracy in healthcare

3. Methods

This scoping review aims to explore and map the use of *LLM*s in the field of medicine, focusing on three main aspects: temporal analysis, geographical analysis, and collaboration networks. The objective is to provide a comprehensive overview of the evolution, global trends, and international cooperation in this emerging field, without delving into a detailed quantitative synthesis of the results.

The review follows a structured, multiphase methodology, as illustrated in Figure 2, encompassing data collection, analysis, and interpretation. The methodology adheres to established guidelines for scoping reviews to ensure clarity, transparency, and reproducibility.

The first two are essential for data collection, while phase 3 may include various types of analyses in order to study some bibliometric indicator. Finally, the last step involves a discussion of the results with the goal of explaining different behaviors.

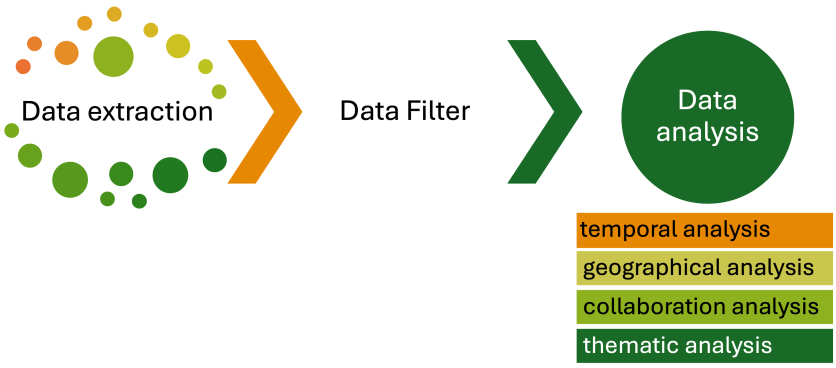


Figure 2. Processing flow of this systematic review.

3.1. Data Collection

The research will follow a structured methodology to identify, collect, and map relevant studies on the use of *LLM*s in medicine—it will include peer-reviewed articles, research reports, theses, conference proceedings, and corporate documents that discuss the use of *LLM*s in the medical context, with particular attention to describing the architectural models employed, application areas, temporal trends, and international collaborations.

The study incorporates a comprehensive search across scientific databases (PubMed, Scopus, IEEE Xplore, and Google Scholar) to capture peer-reviewed articles, research

reports, theses, conference proceedings, and corporate documents discussing the use of *LLM*s in medicine. The search strategy is designed to maximize coverage and relevance by using a combination of keywords related to large language models, artificial intelligence, and healthcare applications. Inclusion and exclusion criteria are defined to focus on works that explicitly mention the architecture, application, and impact of *LLM*s in the medical domain. To minimize bias, the following steps were implemented: (i) Search Strategy Optimization: using broad search terms and synonyms to capture diverse terminologies associated with *LLM*s; (ii) Reviewer Consensus: Multiple reviewers independently screened titles and abstracts to ensure consistent application of inclusion criteria, and discrepancies were resolved through consensus or consultation with a third reviewer; (iii) Data Validation: Cross-verification of selected studies to eliminate duplicates and ensure relevance.

3.2. Bibliometric Indicators

This review uses several bibliometric indicators to analyze temporal trends, geographical distribution, and collaborative networks. We will conduct three types of analyses: temporal-, geographical-, and collaboration-based analyses. Temporal analysis will focus on the evolution of the use of *LLM*s, geographical analysis will focus on the spread of *LLM*s over the world and, finally, the collaboration network analysis will explore interactions between various entities (authors, research centers, or nations) in the adoption of *LLM*s in medicine. To summarize, this paper aims to answer some of the following questions.

- Temporal analysis. To measure the growth rate of research outputs and identify significant technological advancements or pivotal years in *LLM* adoption.
  - When were *LLM*s introduced into the medical field, and how has their application evolved over time?
  - What moments of innovation and technological advancements have driven the adoption of these models in the healthcare sector?
  - How have the areas of application (diagnosis, treatment, medical research, healthcare management, etc.) changed over the years?
  - Have seasonal trends or specific events emerged that accelerated the adoption of *LLM*s in the medical field?
- Geographical analysis. Mapping authors' affiliations to assess regional adoption and the influence of socio-economic factors.
  - In which regions or countries are *LLM*s most widely used in medicine?
  - Are there significant differences in the adoption of *LLM*s in medicine between developed and developing countries?
  - How are *LLM*s utilized in various medical specialties across different geographical areas (e.g., oncology, cardiology, and emergency medicine)?
  - What cultural differences or healthcare policies influence the adoption and use of *LLM*s in medicine?
- Collaboration network analysis.
  - Who are the main actors involved in researching and implementing *LLM*s in medicine?
  - What international partnerships or cross-sector collaborations (e.g., between medical research institutes and technology companies) have emerged in the adoption of *LLM*s?
  - Are there collaborative networks that share resources, data, and knowledge to optimize the use of *LLM*s in the medical field?
  - How do healthcare policies and legal regulations influence the creation and effectiveness of these networks?

- Thematic analysis.
  - What are the primary applications of *LLM*s in the healthcare domain? This question identifies the specific tasks where *LLM*s are applied, such as medical text summarization, clinical decision support, diagnosis prediction, electronic health record (EHR) management, patient communication, and medical research analysis.
  - What are the key benefits of using *LLM*s in healthcare? The analysis highlights the advantages of *LLM*s adoption, including improved efficiency in processing medical data, enhanced diagnostic accuracy, real-time patient support, and reduced administrative workload for healthcare professionals.
  - What challenges and limitations are encountered in the adoption of *LLM*s in healthcare?
  - How do *LLM*s impact patient outcomes and healthcare workflows?
  - What future opportunities and trends can be identified for *LLM*s in health?

In conclusion, this paper is relevant for understanding the adoption and evolution of large language models in the medical field, as it will provide valuable insights on (1) the temporal dynamics and global trends in the adoption of these technologies, (2) regional differences in the implementation of *LLM*s, highlighting areas of significant development and those that could benefit from increased efforts, (3) international and cross-sector collaborations that can foster innovation and the integration of *LLM*s into the healthcare system, and finally, (4) thematic aspects for a comprehensive understanding of the role, challenges, and potential of *LLM*s in revolutionizing healthcare practices.

3.2.1. Temporal Analysis: How Does It Change Over Time?

Temporal analysis allows us to track the evolution of *LLM*s in healthcare, highlighting significant trends and changes. It helps identify how the adoption of *LLM*s has grown or changed over time, pinpointing key moments (e.g., after the introduction of ChatGPT). Moreover, it highlights how academic and clinical interest has increased or shifted, and which areas (diagnosis, training, etc.) have received more attention at different stages. It can demonstrate the impact of new versions of *LLM*s (such as GPT-3 or GPT-4) on their use in healthcare.

In our study, we searched the publications starting from 1995, which can be considered as a watershed, and are divided into three parts, as shown below.

Period 1 (1995–2017): "Pioneering Period" or "Pre-Transformer Era". This period is characterized by by foundational advancements in neural networks and the use of simpler language models, such as rule-based models, Naive Bayes, Hidden Markov Models (HMM), and classical machine learning models. The use of language models was limited to specific tasks, such as information extraction from medical texts and managing electronic health records (EHRs). The proliferation of the internet and large-scale digital content provided vast amounts of text and structured data, essential for training language models. Public datasets like Wikipedia, news corpora, and large repositories of scientific literature became standard training resources. Moreover, the advancements in computational power due to the development and accessibility of Graphics Processing Units (GPUs), followed by specialized hardware like Tensor Processing Units (TPUs), enabled efficient parallel processing, drastically reducing the time and cost required to train deep neural networks. The introduction of Word2Vec (2013) significantly improved the vector representations of words. The introduction of new activation functions (such as ReLU), regularization techniques (e.g., dropout), and advanced optimization algorithms (like Adam) improved model stability and performance. Breakthroughs in unsupervised pretraining (e.g., Word2Vec in 2013 [30]) demonstrated the power of using vast text data to learn meaningful word

representations. IBM Watson (2011) demonstrated the applicability of AI in the medical field, especially in diagnosis. The period witnessed critical steps in evolving neural network designs, with long short-term memory (LSTM) and gated recurrent units (GRU) providing solutions for modeling sequential data. These architectures addressed limitations in processing long-range dependencies in text but still faced challenges in scaling. In 2012, AlexNet's success in the ImageNet competition highlighted the potential of deep learning for image classification, influencing natural language processing (NLP) research. The same momentum carried into NLP with encoder–decoder models, culminating in the transformer architecture (introduced in 2017). Despite these advances, the pre-2017 period was marked by limitations in scaling models. However, it provided a fertile ground for the breakthroughs that followed, particularly the transformer, which replaced recurrence with attention mechanisms and laid the foundation for modern large language models. This period can, thus, be seen as an incubation phase for the *LLM* revolution, driven by continuous progress in data, computing power, and innovative algorithms.

Period 2 (2018–2022): "Period of Transformative Innovation". This period marks the advent of Transformers, such as BERT by Google in 2018 and OpenAI's GPT-2 in 2019, which revolutionized natural language processing. It was a time of accelerated innovation and increasingly widespread adoption in the medical field, and impressive improvement in the ability to process natural language. Usage includes medical text classification, symptom-based diagnosis, and support for scientific research, as well as the implementation of chatbots for healthcare assistance and predictive analytics. Moreover, collaborations between universities emerged, and tech companies (Google Health, IBM, Microsoft) and hospitals developed *LLM*-based solutions. In 2020, GPT-3 demonstrated advanced generative capabilities, and the adoption of models like T5 and Megatron expanded natural language processing capabilities.

Period 3 (2023–Present): "Period of Advanced Application". The current period is characterized by the introduction of advanced models like GPT-4, Claude [31], Mistral, and Gemini [32], which have greatly expanded the ability to handle context and understand complex medical information. Advanced applications have grown, especially in personalized medicine, diagnostics, and clinical decision support systems. *LLM*s have been used to analyze genomes, clinical studies, and personalized treatments, as well as to support clinical decisions. Projects like BLOOM (multilingual) [33] show a collaborative international approach. Finally, the adoption of multimodal models (such as Gemini) that integrate text, medical images (X-rays, CT scans) [34], and clinical data has completed this progress.

3.2.2. Geographical Analysis

This analysis is conducted by extracting information related to the authors' affiliations and associating each paper with the set of countries obtained as the union of all author's current affiliations. Examining the use of *LLM*s in different geographical contexts helps to understand how cultural, economic, and infrastructural factors influence adoption and application. Moreover, it can reveal differences in access and adoption of *LLM*s between high-income and low-income countries, highlighting possible technological or economic barriers. It also shows how regulations and healthcare policies vary globally, impacting the implementation of *LLM*s (e.g., differences between Europe and the USA). The analysis also explores how language barriers or cultural differences influence the effectiveness of *LLM*s (e.g., the adaptation of multilingual models in non-English-speaking contexts). The combination of both temporal and geographical analyses enables the discovery of geographical areas where *LLM*s have spread more rapidly in response to specific events.

3.2.3. Collaboration Analysis

The third analysis centers on collaboration by constructing a network that connects various entities. Each link in the network quantifies the intensity of their collaborative interactions. In this study, the entities involved are: affiliations and countries. More specifically, a collaboration analysis is a graph, where each vertex V_i is an entity and each link (V_i, V_j) between the pair of nodes V_i and V_j represents the fact that they have, at least, one paper in which they both appear. The weight of the link w_{ij}, related to (V_i, V_j), is the number of papers co-authored by V_i and V_j. Let us note that the collaboration relationship is bidirectional, so the graph is undirected.

According to the type of entity represented by V_i we have three different collaboration networks with different granularity:

- The network with the smallest granularity, referred to as *co-authorship networks*, focuses on authors; therefore, V_i represents an author. This type of network is the most commonly used in many applications [35]. The graph obtained is a co-authorship network. A link in this network represents the fact that two researchers are co-authors of at least one paper. This type of collaboration falls outside the scope of this paper.
- The second type of network focus on affiliations and is referred to as *co-affiliation network*. In this case the entity, A_i is a research center such as a department, a laboratory, or a university. In the co-affiliation network, a link exists if (one or more) authors from the two affiliations collaborate on almost a publication. The weight, w_{ij} is the total number of papers which authors belong to A_i and A_j. The density of the graph represents how much the different centers collaborate with each other. The number and size of connected components represent how fragmented the activity is, the degree of a node represents the number of collaborations, and the ratio between degree and weighted degree indicates the frequency of collaborations between centers A_i and A_j.
- The third type of entity is the country, referred to as *co-nationalities network*; therefore, an entity represents a country, C_i. Note that we will refer to the country where the affiliation of a researcher is located, not the nationality of the researcher. In this case, the link (N_i, N_j) exists if there is a paper whose co-authors come from affiliations located in countries N_i and N_j. The Weight degree of the link is the intensity of the collaboration between the two countries, while the degree of N_i represents how many countries N_i collaborates with. Finally, in this case, the ratio between the weighted degree and the degree measures the average intensity of collaborations of country N_i.

4. Data Collection on *LLM* Literature in Medicine

For our dataset, we focused on technological advancements and their applications in medicine, referencing the Scopus and PubMed databases. The structured organization of Scopus was utilized to filter the publications included in our analysis. Scopus indexes papers with a set of author-defined keywords, enabling targeted filtering. Publications were selected if their title, abstract, or keywords contained at least one term from each column in Table 2.

Acknowledging the recent emergence of this topic, the analysis was restricted to papers published from 1995 onwards. This filtering process identified a total of 14,952 papers, of which 4899 were also indexed in PubMed, while the remaining 10,053 were unique to Scopus.

Table 2. The query is composed of two terms, the first referring to AI model and the second referring to Medicine. Both must be present.

Condition A	Condition B
large language model	Medicine Applications
LLM	Healthcare
Transformer Model	Medical Application
GPT	Clinical Application
BERT	Diagnosis
	Treatment
	Health Informatics
	Medical AI
	Clinical Decision Support
	Medical Assistant
	Drug Discovery
	Biomedical Research

To ensure a more comprehensive dataset, the decision was made to include all papers available on Scopus, including those not indexed in PubMed, extracting all information described in Table 3.

Table 3. Fields extracted from Scopus and/or PubMed. The third column indicates if the parameters has been searched for containing the keys.

Parameter	Description	Searched for
Authors	contains the list of authors	Yes
Title	contain the title of the paper	Yes
Abstract	contains the text of the abstract	Yes
Key	contains the list of the keyword provided from the authors	Yes
Affiliation	contains the list of the affiliation involved in the pubblication	
Country	contains the list of the affiliation country	
Citation	contains the number of the citation of the paper, this value are referred to citation on Scopus Year of pupublication	
Pubmed	is a Boolean; it is True if the paper is also indexed by Pubmed	
Type	it refers to the publication type. e.g., “review”, “conference paper”	

This dataset will be referred to throughout the rest of the paper as *All*. Figure 3 shows the number of publications per year from the extracted dataset.

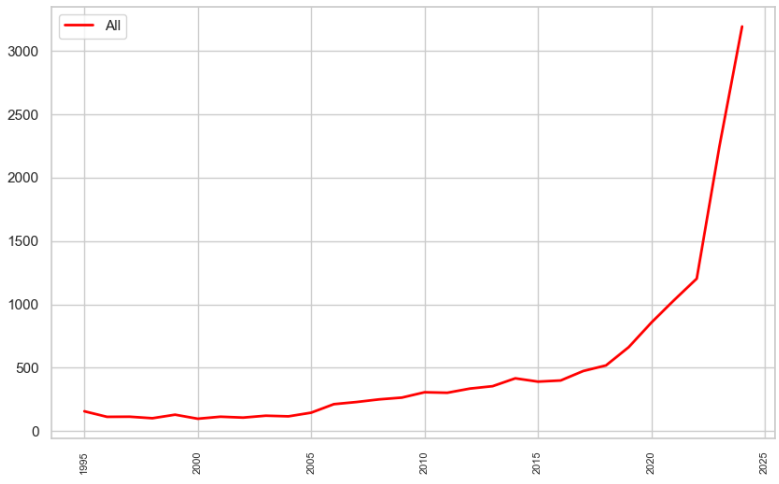


Figure 3. Distribution of papers over the years.

As illustrated in Figure 3, the number of publications exhibits significant variation across different periods, with notable peaks in 2017 and 2021. Based on this observation, the dataset has been segmented into three distinct periods, designated as *Early*, *Medium*, and *New*, corresponding to All_{Early}, All_{Medium}, and All_{New}, respectively.

Additionally, to conduct a more detailed analysis of the literature, we extracted the publications which type is "Review" in order to compare them with the global corpus of publications. This division leads to two datasets: *All*, which contains all the extracted publications, and *Review*, which contains only reviews. Table 4 and provided some information about these datasets. Furthermore, each of these datasets has been divided into Early, Medium, and New. In the table, the sizes of the individual datasets and the number of publications per year in each examined period are shown.

Table 4. Dataset facts.

DB	Period	#Papers	#Papers Yearly
All	1995–2024	14,952	498.4
All_{Early}	1995–2017	5240	227.8
All_{Medium}	2018–2021	3069	762.25
All_{New}	2022–2024	6643	2214.4
Review	1995–2024	483	16.1
$Review_{Early}$	1995–2017	86	3.73
$Review_{Medium}$	2018–2021	46	11.5
$Review_{New}$	2022–2024	350	117.7

For both datasets, it can be observed that the number of publications per year increases significantly as we move forward across the three periods. The two datasets, *All* and *Review*, will be used for our quantitative analysis, and for each of them, we will also analyze the databases covering specific periods. The analysis conducted in this section is purely quantitative, aiming to highlight trends in publications on the topic without focusing on individual papers or the specific themes addressed in each work.

In the last part of the paper, we aim to perform a Thematic Analysis of the Literature on the most-cited papers. To achieve this, given the large size of the previous datasets, we applied a filtering process to extract the most significant papers from the two datasets.

Determining the "most significant" paper is inherently subjective and open to extensive debate. While citation counts are a commonly used quantitative metric to gauge a paper's impact, they have notable limitations. Citation counts measure impact, not quality, and can be influenced by factors such as academic trends, self-citations, and varying citation practices across disciplines. However, for the sake of simplicity and given the scope of this review, we evaluated the importance of papers using the number of citations they have accrued (with citation counts obtained from Scopus). Since this metric is heavily influenced by the publication year of the paper, we computed the number of citations per year. Papers were selected if their citations-per-year value exceeded four times the average citations-per-year value. From this filtering process, we constructed the datasets $Dataset_{Main}$ and $Dataset_{MainReview}$.

In Table 5, some characteristics of these datasets are shown. Also in this case, the number of papers (#Papers) per year increases in the section containing the most recent papers. Similarly, in these two cases, the number of papers is much higher in the most recent period.

Note that the number of reviews satisfying the "most important" criteria, as defined, is very small. Since the number of papers belonging to the most cited reviews is limited, a statistical analysis is not meaningful. Therefore, we provide a brief summary of all of them,

highlighting their focus and achievements, if any. Since we aim to exploit the emerging topics, we refer only to $MainReview_{New}$ and $MainReview_{New}$.

Table 5. Main statistics for the databases Main and *MainReview*.

Database	Period	#Papers	#Papers Yearly
Main	1995–2024	636	21.2
$Main_{Early}$	1995–2017	227	9.87
$Main_{Medium}$	2018–2021	138	33.5
$Main_{New}$	2022–2024	287	95.67
MainReview	1995–2024	20	0.67
$MainReview_{Early}$	1995–2017	6	0.26
$MainReview_{Mediun}$	2018–2021	2	0.5
$MainReview_{New}$	2022–2024	14	4.67

5. Results: Quantitative Analysis and Most Important Publications

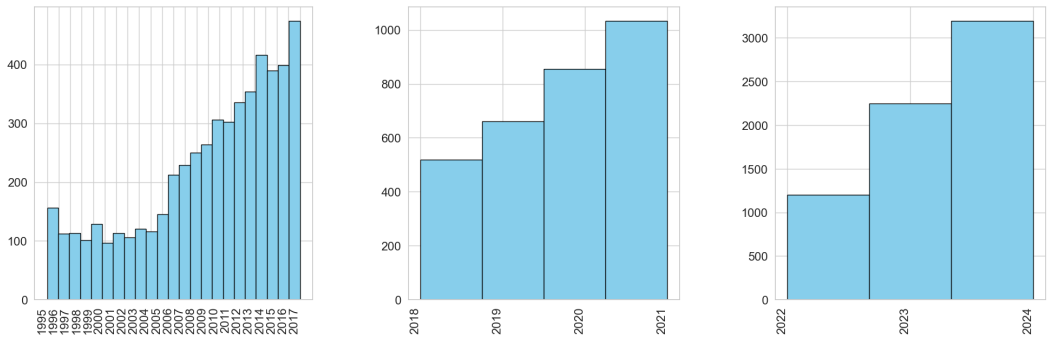
This section provides a comprehensive summary of all the results obtained throughout the analysis. The first three subsections focus on the quantitative analysis, presenting detailed data and statistical findings. In contrast, the following two subsections are dedicated to the analysis of the main papers, offering a deeper understanding of their content and significance within the context of the study

5.1. Temporal Analysis

This section present the temporal analysis on all datasets discussed in the above section. Both trends and notes will be discussed in detail.

Figure 4 shows the trend in the number of publications across the three periods. Pay attention to the scale used. The y-axis in Figure 4a is 500, in Figure 4b it is 1000, and in Figure 4c it is 3500. This demonstrates an increasingly rapid growth in the number of publications over time. The number of publications per year has tripled from one period to the next. The trend within each period is almost always upward, except for a few sporadic years in the first period. This indicates a growing interest in the applications of *LLM*s in medicine.

The same analysis conducted on the *Review* dataset reveals a growth in the number of papers per year by an order of magnitude when moving from $Review_{Medium}$ to $Review_{New}$. The strong demand for "reviews" may be significant because reviews play a key role in unifying approaches from different disciplines, fostering collaboration and knowledge exchange. In particular, the increased use of *LLM*s in medicine in recent years may have caused this surge in reviews (see Table 4).



(**a**) All_{Early} (**b**) All_{Medium} (**c**) All_{New}

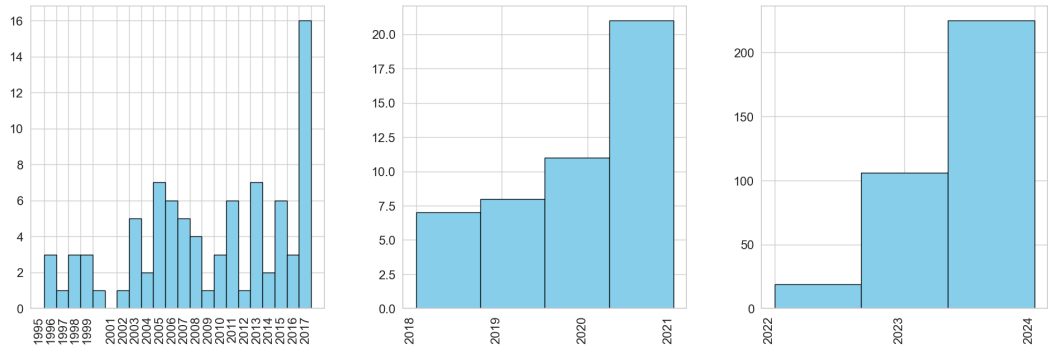
Figure 4. Distribution of the *All* dataset.

Figure 4 illustrates publication count trend while Figure 5 deals with publications classified as "Review".

During the first period, shown in Figure 5a, the number of review publications remained below six per year, with the notable exception of 2017, which experienced a sudden surge to 16 publications. This indicates that 2017 marks the initial significant increase in interest surrounding *LLM*s.

Figure 5b highlights the trend in the intermediate period, beginning with 7 reviews in 2018 and exhibiting a steady rise, reaching 21 publications by the end of this phase.

A more pronounced shift is observed in the final period, as shown in Figure 5c. Likely driven by the release of products like ChatGPT, the number of annual review publications reflects a significant growth in interest, climbing from approximately 20 papers to over 200 in 2024. This underscores the accelerating engagement with and practical relevance of *LLM*s in recent years.



(**a**) $Review_{Early}$ (**b**) $Review_{Medium}$ (**c**) $Review_{New}$

Figure 5. Distribution of the *Review* datasets.

It should be noted that the data for 2024 are not complete, as the data extraction was performed in November 2024.

5.2. Geographical Analysis Results

This section presents the geographical analysis focused on attention of each country to the technologies related to *LLM*s and their applications in health.

To facilitate this analysis, each paper was linked to one or more countries based on the affiliations of the authors, as provided in Scopus. When multiple authors of the same paper were affiliated with the same research center or university, the corresponding country was included only once in the list of affiliations for that paper.

The contribution of each country was then quantified by aggregating the number of papers where the country appeared in the affiliation list. A higher count indicates a greater level of research activity and interest in the topic.

Figure 6 presents a pie chart depicting the 15 countries with the highest research impact, collectively accounting for approximately 25% of the total contributions. The remaining countries are consolidated into a single category labeled as "Others".

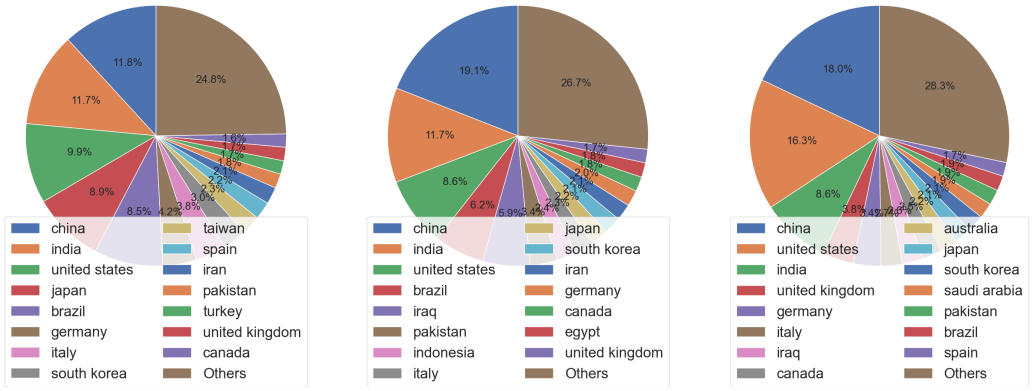
From these charts, it is evident that China, India, and the United States contribute most significantly (though in varying order) and collectively account for about 30% in the first period (see Figure 6a) and approximately 40% in more recent periods (Figure 6b,c).

Some countries, such as the United Kingdom, Germany, Brazil, Italy, and South Korea, are consistently present across all periods. Other countries, like Japan and Taiwan, appear in the earlier periods but are no longer present in recent years.

While it might seem surprising to find less developed countries in this ranking, scientific research is becoming increasingly global, with researchers from diverse nations

collaborating. Despite advancements, a digital divide continues to separate developed and developing countries. The intense competition for leadership in this field is evident in government policies, investments, and the movement of skilled workers. Nevertheless, the participation of less developed countries highlights AI's potential to drive development and bridge gaps.

Finally, new countries emerge in the most recent period: Iraq and Saudi Arabia, reflecting a growing interest in *LLM* technologies applied to medicine across a broader range of regions.



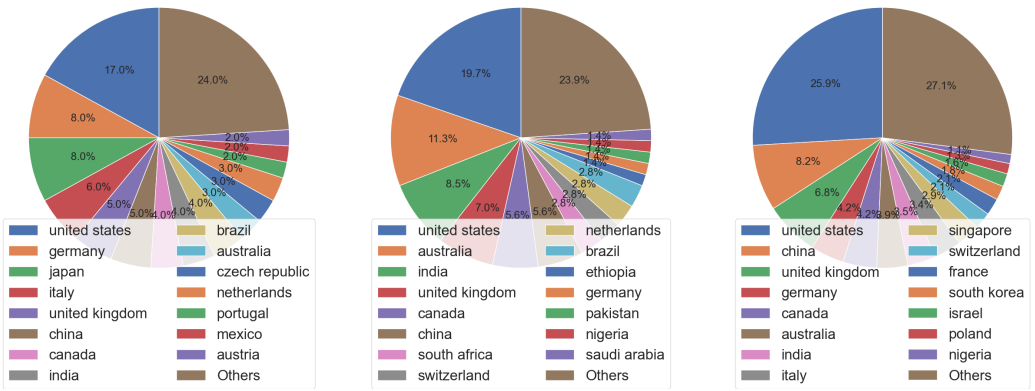
(**a**) All_{Early} (**b**) All_{Medium} (**c**) All_{New}

Figure 6. The 15 countries with the highest research impact for *All*.

Performing the same analysis on "Review" publications reveals that the primary interest remains centered in the same countries, as shown in Figure 7. Since authors of review articles often aim to disseminate and popularize new research topics, this is likely correlated with the emergence of new areas of interest in *LLM*s in medicine in different countries.

In the first period (see Figure 6a), the United States dominates with 17% of the publications. Besides the United States, notable contributors include Germany (8%), Italy (6%), and the United Kingdom (5%). Asia is represented by China (5%) and India (4%). During this period, the landscape of publications appears relatively distributed, with limited concentrations of dominance and significant contributions from European countries.

The main changes in the intermediate period (Figure 6b) include an increase in contributions from the United States, while India (8.5%) and Australia (7%) become more prominent actors. New contributors emerge, such as Saudi Arabia, Nigeria, and Ethiopia, reflecting a global expansion of interest in *LLM*s applied to medicine.



(**a**) $Review_{Early}$ (**b**) $Review_{Medium}$ (**c**) $Review_{New}$

Figure 7. The 15 countries with the highest research impact in review-type publications.

In the most recent period (Figure 6c), the United States further increased its contribution, reaching 25.9%. China became the second most significant contributor with 8.2%, surpassing India and Germany. Additionally, in this recent period, new countries such as Nigeria, Poland, and Israel appear, showing an increasing interest in publishing review articles.

In conclusion, the countries most actively engaged in this area of research are often the same; however, some variations can be observed in recent periods, both in general across all types of papers and specifically in those of the "Review" type.

5.3. Collaboration Analysis

This section presents an analysis of collaboration networks, both in terms of collaborations between affiliations and countries. This study is conducted for both the *All* dataset and its partitions All_{Early}, All_{Medium}, and All_{New}, as well as for the *Review* dataset and its partitions $Review_{Early}$, $Review_{Medium}$, and $Review_{New}$.

Table 6 shows the characteristics of the collaboration networks between affiliations (ACN_i) in the first four rows and those of the collaboration networks between countries (ACN_i) for the *All* dataset. For each network, the table reports the network size in terms of node and edge cardinality ($|V|$ and $|E|$, respectively), the maximum and average degree values (maxD and $\overline{D}$, respectively), the maximum and average weighted degree values (maxWD and $\overline{WD}$, respectively), the average ratio between degree and weighted degree $\overline{D/WD}$, and the network density. As expected, the size of the collaboration networks is significantly larger than that of the country-level collaboration networks, although the latter exhibit much higher density ρ.

The networks show similar maxD values, whereas $\overline{D}$ values are much higher in country-level collaboration networks. Regarding weighted degree, both the maximum and average values are much higher in the country-level collaboration networks. Finally, the $\overline{D/WD}$ ratio is higher in affiliation-level collaboration networks. Overall, the highest collaboration values are consistently observed in the most recent periods.

Table 6. Affiliation collaboration network (ACN) and nationalities collaboration network (NCN) characteristics.

	Network	$\|V\|$	$\|E\|$	$maxD$	$\overline{D}$	$maxWD$	$\overline{WD}$	$\overline{(D/WD)}$	ρ
ACN	ACN_{all}	38,423	73,135	132	3.81	138.00	3.94	0.89	0.0001
	ACN_{early}	12,125	18,399	57	3.03	112.00	3.23	0.85	0.0003
	ACN_{medium}	7724	12,623	44	3.27	54.00	3.33	0.89	0.0004
	ACN_{new}	18,994	42,214	119	4.44	123.00	4.56	0.91	0.0002
NCN	NCN_{all}	150	1749	99	23.32	1556.00	89.65	0.60	0.1565
	NCN_{early}	107	563	57	10.52	326.00	23.85	0.67	0.0993
	NCN_{medium}	109	492	52	9.03	212.00	18.39	0.70	0.0836
	NCN_{new}	138	1445	88	20.94	1018.00	64.43	0.63	0.1529

Table 7 shows, in the first part, the affiliations with the maximum degree and, in the second part, the countries with the maximum degree in each network. This indicates the affiliations (or countries) with the highest number of connected affiliations (or countries) as collaborators. The third column lists those with the maximum weighted degree (WDegree), representing the entities with the highest total number of collaborations.

MaxDegree and MaxWDegree can differ significantly. For example, an affiliation (or country) may have one collaborator with whom it has a high number of interactions, while simultaneously having many collaborators with whom it has few connections.

Table 7. Affiliation with maxD in ACN and NCN built on *All*.

Network	*maxD*
ACN_{all}	Harvard Medical School Boston MA United States
ACN_{early}	Department of Pediatrics Yamato City Hospital Japan
ACN_{medium}	Department of Psychiatry and Psychotherapy University Medical Center Mainz Mainz Germany
ACN_{new}	Harvard Medical School Boston MA United States
NCN_{all}	United States
NCN_{early}	United States
NCN_{medium}	United States
NCN_{new}	United States

Tables 8 present the main characteristics of the ACN and NCN networks built on the *Review* dataset. As in the case of *All*, the tables are organized into two sections: the first four rows provide information regarding the collaboration networks among affiliations (ACN), while the subsequent four rows contain information about the collaboration networks among countries (NCN).

The network size ($|V|$ and $|E|$) grows significantly in the most recent period. The network density is higher in the early and intermediate periods compared to the entire period and the most recent period. This suggests that some new affiliations have joined the network but maintain a limited number of collaborations. The average values of avgD and avgWD progressively increase, reaching their highest levels in the most recent period (7.58 and 7.74, respectively).

The collaboration networks among countries (NCN) are, as expected, smaller compared to those among affiliations (ACN) but exhibit higher densities. The maximum and average weighted degree values (maxWD and avgWD) are notably high in the most recent period (180.00 and 16.27, respectively). In the recent period, the number of nodes $|V|$ and edges $|E|$ increases significantly, indicating an expansion of the network of collaborations among nations. The density of the NCN_{newR} network (0.1329) is the highest across all periods, highlighting a more connected and centralized collaboration network.

Overall, the tables demonstrate that collaborations, both at the level of affiliations and nations, show significant growth in the most recent period, with larger networks and increasing values for both average degree and weighted degree.

Table 9 lists the affiliations and countries with maximum degree in ACN and NCN networks built on *Review*. It is worth noting that, among the affiliations, entities from Singapore and Hong Kong are present, while for countries, the United States in the most recent period is supplanted by the United Kingdom. All the affiliations are involved in health.

Table 8. Affiliation collaboration network (ACN) and nationalities collaboration network (NCN) characteristics extracted from review papers.

Network	$\|V\|$	$\|E\|$	*maxD*	$\overline{D}$	*maxWD*	$\overline{WD}$	$\overline{(D/WD)}$	ρ
ACN_{allR}	1870	6635	57	7.10	60.00	7.25	0.93	0.0038
ACN_{earlyR}	193	276	10	2.86	10.00	2.86	0.81	0.0149
$ACN_{mediumR}$	172	635	27	7.38	27.00	7.45	0.91	0.0432
ACN_{newR}	1514	5737	57	7.58	60.00	7.74	0.95	0.0050
NCN_{allR}	86	445	59	10.35	203.00	17.33	0.71	0.1218
NCN_{earlyR}	32	50	12	3.12	12.00	3.19	0.68	0.1008
$NCN_{mediumR}$	33	31	7	1.88	13.00	2.61	0.48	0.0587
NCN_{newR}	80	420	58	10.50	180.00	16.27	0.76	0.1329

Table 9. Affiliation and country with maximum degree values on a collaboration network built on *Review*.

Network	*maxD*
ACN_{all}	Singapore Eye Research Institute Singapore National Eye Centre Singapore
ACN_{early}	Children's Hospital of New York Columbia University New York NY United States
ACN_{medium}	Accident and Emergency Medicine Chinese University of Hong Kong Faculty of Medicine Hong Kong
ACN_{new}	Singapore Eye Research Institute Singapore National Eye Centre Singapore
NCN_{allR}	United States
NCN_{earlyR}	United States
$NCN_{mediumR}$	United Kingdom
NCN_{newR}	United Kingdom

To better understand the collaboration networks, graphical representations were created specifically for NCNs. Representations for ACNs were not included, as their extensive size and high number of nodes make them less descriptive. Instead, NCN subgraphs were drawn focusing on the 15 countries with the highest degree, highlighting key features of the collaborations.

In Figure 8, the network constructed from all reviews, restricted to the 15 countries with the highest degree, is displayed. As can be seen, this network highlights certain countries with higher collaboration intensity (the links in the figure are weighted proportionally to their strength). Shown also in the figure are the 20 strongest links. In particular, collaborations between the United States and Canada, China, Germany, Italy, and the United Kingdom are especially strong, with the United States serving as the central hub of these collaborations. The structure of this figure is the same as that found in the subsequent figures.

Examining collaborations over time reveals that, in the network built from All_{Early} (see Figure 9, representing the oldest papers), the central core also includes Japan and Taiwan. This suggests that some collaborations present during the earlier periods have gradually diminished over time. The most significant collaborations are those between the United States and China, but in general, the United States is involved in most of the strongest links. A noteworthy collaboration between China and Taiwan should also be highlighted.

Country 1	Country 2	Weight
China	United States	184
Canada	United States	115
United Kingdom	United States	98
Germany	United States	86
Pakistan	China	79
United States	India	69
United States	Italy	66
Japan	China	51
China	United Kingdom	49
Brazil	United States	45
United Kingdom	Italy	45
United States	South Korea	45
Japan	United States	42
Australia	United States	41
Australia	China	38
Germany	United Kingdom	34
United Kingdom	India	33
China	India	31
Australia	United Kingdom	29
United Kingdom	Canada	28

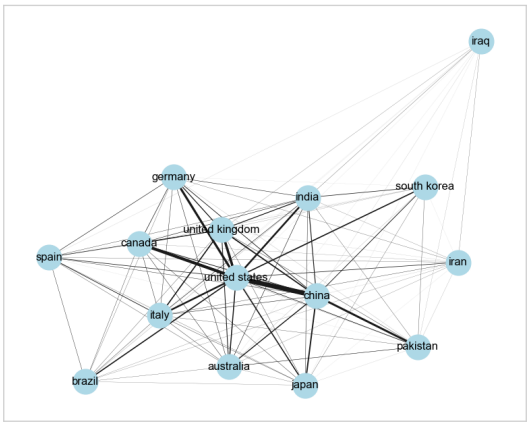


Figure 8. The 15 countries with the highest degree in NCN_{all}. The weights, reported in the table, are proportional to the thickness of the links in the network.

In the intermediate period, Figure 10, the situation is significantly different. Two links are particularly strong, indicating more intense collaboration between China and the United

States, and between China and Pakistan. The set of nations in this network is quite distinct from the two previous cases.

Country 1	Country 2	Weight
China	United States	32
China	Taiwan	31
Canada	United States	26
South Korea	United States	20
Japan	China	18
United States	Japan	18
United Kingdom	Italy	17
Germany	United States	16
United States	United Kingdom	15
United States	Italy	16
Brazil	United States	15
United States	India	11
Germany	China	9
United States	Turkey	8
Germany	India	7
Spain	Italy	7
China	South Korea	7
Germany	United Kingdom	6
Japan	Italy	6
China	Canada	6

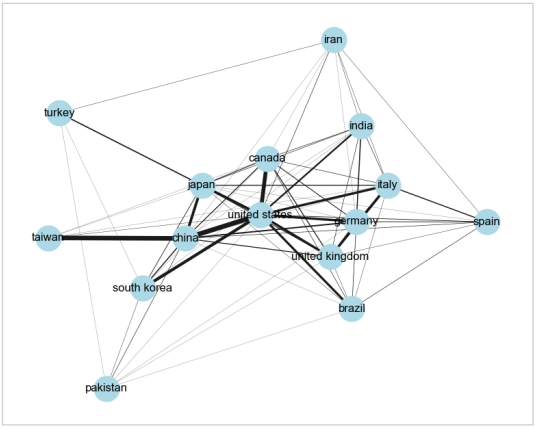


Figure 9. The 15 countries with the highest degree in NCN_{EarlyR}. The weights, reported in the table, are proportional to the thickness of the links in the network.

Country 1	Country 2	Weight
China	United States	36
Pakistan	China	31
Canada	United states	22
Germany	United states	12
United States	India	12
China	India	11
India	South Korea	11
United Kingdom	United States	8
Japan	China	7
Brazil	United States	7
United States	Italy	7
China	South Korea	6
Egypt	China	5
Germany	Canada	5
Japan	United States	5
China	United Kingdom	5
United Kingdom	India	5
Pakistan	United States	4
Iran	Italy	4
Canada	United Kingdom	4

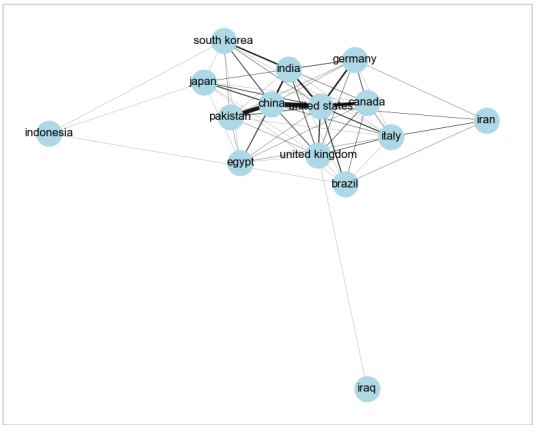


Figure 10. The 15 countries with the highest degree in NCN_{Medium}. The weights, reported in the table, are proportional to the thickness of the links in the network.

Finally, the collaboration network for the most recent period (see Figure 11) shows a network with more heavily weighted links centered around the United States, involving many other countries with a worldwide distribution. The United States emerges as the central hub, with significant collaborations with Europe, Asia, and other regions. European countries (United Kingdom, Germany, and Italy) are well represented with a total of six links, often involving the United States or collaborating among themselves (e.g., Germany and the United Kingdom, the United Kingdom and Italy). China is a key node with six links, primarily with other Asian countries (Japan, Pakistan, Saudi Arabia) and major powers such as the United States and the United Kingdom. Saudi Arabia stands out as a notable partner with three links, two of which are with other Asian countries (Pakistan and China). No African links are reported, suggesting marginalization in this network.

Figure 12 displays the network built from the entire *Review* database. The countries with the highest levels of collaboration differ from those in the case of all papers, indicating that the collaborations of certain countries are more specifically directed toward the publication of review papers on the topic. In addition to the countries observed in the general database, nations such as India and Israel also exhibit higher levels of collaboration.

Nevertheless, the countries with the strongest collaborations remain the United States, the United Kingdom, and China.

Country 1	Country 2	Weight
China	United States	116
United Kingdom	United States	74
Canada	United States	67
Germany	United States	58
United States	India	46
Pakistan	China	45
Pakistan	Saudi Arabia	43
United States	italy	43
Saudi Arabia	india	42
China	United Kingdom	38
Australia	United States	34
Australia	China	30
China	Saudi Arabia	29
Japan	China	26
Germany	United kingdom	25
United Kingdom	India	25
Australia	United Kingdom	24
United kingdom	Italy	24
Brazil	United States	23
United States	South Korea	22

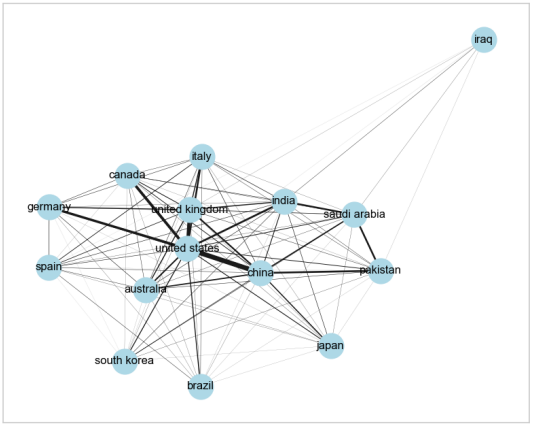


Figure 11. The 15 countries with the highest degree in NCN_{New}. The weights, reported in the table, are proportional to the thickness of the links in the network.

Country 1	Country 2	Weight
United Kingdom	United States	28
China	United States	15
Canada	United States	15
United Kingdom	canada	12
Israel	United States	10
Australia	United States	9
United Kingdom	India	9
United States	Switzerland	9
Australia	United Kingdom	8
Singapore	United States	8
Germany	United States	8
United States	Italy	8
Singapore	United kingdom	7
Germany	United kingdom	6
Germany	switzerland	6
Netherlands	United States	5
United kingdom	Switzerland	5
France	United States	5
Netherlands	United Kingdom	4
Australia	Canada	4

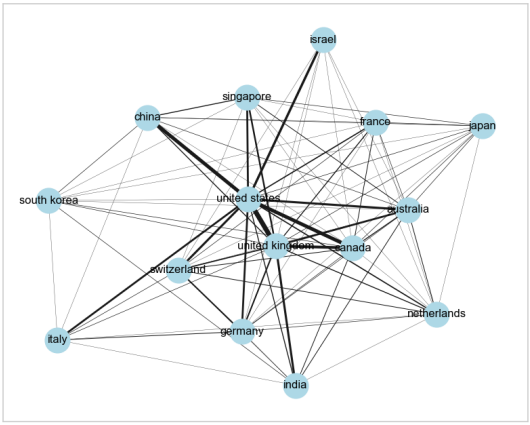


Figure 12. The 15 countries with the highest degree in NCN_{allR}. The weights, reported in the table, are proportional to the thickness of the links in the network.

Looking at the data, we observe that in the periods labeled as "Early" and "Medium", the number of links is extremely small. Consequently, network analysis during these periods is not particularly meaningful.

In the most recent period, (see Figure 13), where the number of review-type publications is larger, the graph becomes fully connected again. The United States emerges as the most connected node with the highest edge weights, highlighting its central position in the network and its intensive collaborations with numerous countries. The United Kingdom is also a central node, maintaining strong ties with the United States (17), Canada (9), and a moderate connection with Singapore (7). Countries such as the United States, the United Kingdom, Canada, Australia, and Singapore exhibit strong mutual connections, reflecting historical, cultural, and political alliances (e.g., Commonwealth ties or technological and military collaborations). The U.S.–China axis: Despite political tensions, the high weight (14) suggests a significant relationship, likely in economic, technological, or scientific domains. Several European countries, such as Germany, Italy, Switzerland, and France, maintain relevant collaborations with the United States, though with lower weights compared to the "Anglosphere". Singapore plays an intriguing role, collaborating with the

United States (8), the United Kingdom (7), and China (4), suggesting a strategic position as an Asian hub for global connections.

The dominant role of the United States, serving as the main hub in the network, connected to nearly all countries with the highest weights, reflects its global influence in fields such as science, technology, economics, and geopolitics.

Country 1	Country 2	Weight
United Kingdom	United States	17
China	United States	14
Canada	United States	12
United Kingdom	Canada	9
Israel	United States	8
Australia	United States	8
Singapore	United States	8
Singapore	United Kingdom	7
Germany	United States	7
United States	Italy	7
United Kingdom	India	6
United States	Switzerland	6
Australia	United Kingdom	5
Germany	United Kingdom	5
Germany	Switzerland	5
Poland	United States	4
Poland	China	4
Singapore	China	4
China	United Kingdom	4
France	United States	4

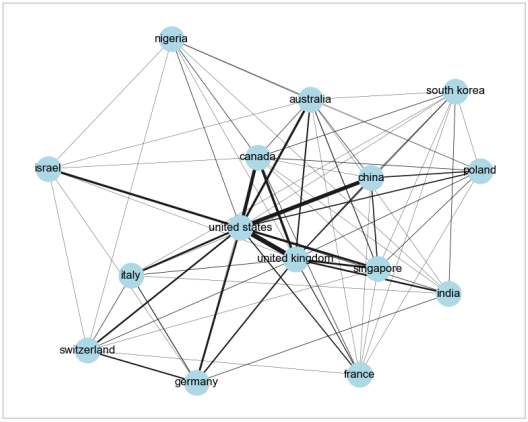


Figure 13. The 15 countries with the highest degree in NCN_{NewR}. The weights, reported in the table, are proportional to the thickness of the links in the network.

5.4. Keyword Analysis of $Main_{new}$

This analysis aims to examine the topics of greatest interest using the $Main_{new}$ dataset; in particular, to discover emerging topics using several AI and NLP tools.

The analysis was conducted using the keywords provided by the authors to describe their articles (the field "author keywords" in the BibTeX file). This keyword list is extremely broad due to the freedom authors have in attributing keywords to each article. Therefore, these keywords were processed using various NLP techniques to reduce redundancy by grouping similar or synonymous keywords. In addition to basic operations, such as case unification and stemming, we performed keyword similarity analysis and synonym analysis using NLTK [36] with a threshold of 0.7.

The number of keywords obtained is approximately 500. In Table 10, the most frequent keywords are shown.

The three clusters are as follows: **I** refer to AI in general, **L** refers to techniques related to *LLM*s, while **M** corresponds to those related to medical applications. Note that the majority of keywords fall under cluster **I**. In cluster **L**, there is a clear prevalence of ChatGPT in its various versions. In cluster **M**, there is a predominance of general keywords that do not refer to a specific pathology or branch of medicine. The keywords associated with cluster **M** are numerous, but all have a low incidence.

Figure 14 shows the cloud art for the keywords from 2022 to 2024. As observed, the incidence of keywords changes over the past three years. Notably, only in 2022, the keywords "Bert" and "ChatGPT" were not dominant, with other keywords such as "Deep Learning" and "Visual Transformer" being more prominent. In 2023, the most relevant keyword was "Artificial Intelligence", followed by "ChatGPT". However, in 2024, the trend reverses, with "ChatGPT" emerging as the most used keyword, reflecting the widespread adoption of this technology in recent times.

(**a**) 2022 (**b**) 2023 (**c**) 2024

Figure 14. Cloud art keyword over years.

Table 10. The most frequent keywords.

Key	Ntot	Cluster
chatgpt	117	L
artificial intelligence	112	I
large learning models	64	L
natural language processing	38	I
vision transformer	34	I
deep learning	27	I
large language models	23	L
gpt-4	23	L
machine learning	20	I
generative pre-trained transformer	17	L
medical education	17	M
healthcare	16	M
fault diagnosis	15	M
pathology	14	M
clinical decision support	13	M
generative ai	13	I
medicine	13	M
google bard	10	L
generative language models	9	L
chatbots	8	I
medical image analysis	8	M
prompt engineering	8	I
communication	7	I
digital pathology	7	M
radiology	7	M
conversational agents	6	L
maize disease classification	6	M
treatment	6	M
antioxidant	5	M
clinical applications	5	M
diagnostic medicine	5	M
disinformation	5	M
ethics	5	M
frcr examination	5	M
general practice	5	M
higher education	5	M
3d segmentation	4	I
accuracy	4	I
bert	4	L
clinical named entity recognition	4	M
clinical settings	4	M
data privacy	4	I
depression	4	M
foundational model (fms)	4	I
mechanical properties	4	M
medical exams	4	M
otolaryngology	4	M
patient education material	4	M

5.5. A Deeper Discussion of Papers in $MainReview_{New}$

Given the limited number of papers included in $MainReview_{New}$, we provide a comprehensive discussion of each. Table 11 reports the month of publication and the thematic areas of the publishing journal, as classified by Scimago Research Centers, which maintain the widely recognized Journal & Country Rank (SJR),

Kanbach et al. [37] reviewed the transformative impact of generative algorithms on the implementation of artificial intelligence (AI) across academia, business, and broader societal contexts. The paper highlights that the advent of generative AI tools, such as ChatGPT, Jasper, and DALL-E, represents a significant milestone in accelerating AI adoption, attributing this progress to their user-friendly design, intuitive interfaces, and advanced performance capabilities. The study adopts a Business Model Innovation perspective to analyze generative AI tools, offering two main contributions: (1) the formulation of six comprehensive propositions elucidating GAI's impact on business practices and (2) an examination of three industry use cases, specifically software engineering, healthcare, and financial services. The scoping review methodology supports the qualitative content analysis, leveraging a broad dataset comprising 513 data points.

Ref. [10] reviewed studies published following the release of ChatGPT, with a particular emphasis on its adoption among the general public and medical professionals. It highlights the "productization" of advanced technologies, enabling the accessibility of artificial intelligence (AI) tools to non-technical users, while stressing the critical precautions necessary for healthcare researchers. The authors conducted a systematic review of publications addressing ChatGPT's applications in healthcare, offering perspectives targeted at general audiences, medical practitioners, and NLP researchers. Articles were retrieved from the PubMed database using the keyword "ChatGPT", with inclusion criteria and a structured taxonomy employed to filter and organize the results. The study concludes that domain-specific NLP models trained on biomedical datasets continue to represent the preferred choice for critical clinical applications.

The authors propose a two-dimensional taxonomy focusing on application domains and user types, categorizing the reviewed publications accordingly. Additionally, a tagging system (Level 1 to Level 3) is applied to denote the depth and specificity of each work.

In [10], the authors analyzed the multifaceted applications of large language models (*LLM*s), such as ChatGPT, within the domains of biomedicine and healthcare. The study focuses on tasks including biomedical information retrieval, question answering, medical text summarization, information extraction, and medical education, evaluating whether *LLM*s possess the transformative potential to revolutionize these fields or whether the intrinsic complexities of the biomedical domain pose distinctive challenges.

The authors conduct a comprehensive literature review, concluding that while *LLM*s have surpassed previous state-of-the-art methods in text generation tasks, their advancements in other applications remain incremental. Furthermore, the study identifies several risks and challenges associated with employing *LLM*s in biomedicine and healthcare, such as the generation of fabricated information in model outputs and significant legal and privacy concerns linked to the handling of sensitive patient data.

The study proposed by Eggmann et al. [38] belongs to a different context since it qualitatively discuss the impact of AI and seems focused on medical matters rather than technical or information. This paper does not provide a systematic review of publications but discusses the impact of large language model on a single specific medicine field such as dental medicine. Moreover, the approach is not extensive, but it discusses the main problems related with medicine matters. The authors concluded that safeguards are essential to mitigate risks like providing outdated, biased, or misleading health information. Additionally, patient confidentiality and cybersecurity concerns must be addressed. Regarding the

authors' "clinical significance" (sic.), they concluded that the adoption of *LLM*s in dental medicine offers supplementary tools for enhancing various aspects of the field but requires thorough consideration of their inherent risks and limitations to ensure safe and effective implementation.

The paper by Liu et al. [39] provides a comprehensive survey of research on ChatGPT, specifically focusing on GPT-3.5 and GPT-4, with an emphasis on their applications. The study discusses key advancements, such as large-scale pre-training, instruction fine-tuning, and Reinforcement Learning from Human Feedback (RLHF), which have notably enhanced the adaptability and effectiveness of large language models (*LLM*s). By analyzing 194 relevant papers from arXiv, the research offers trend evaluations, word cloud visualizations, and domain-specific distribution analyses. The findings highlight a growing interest in ChatGPT-related studies, especially within natural language processing, and uncover promising applications in various fields including education, medicine, mathematics, and physics.

The paper by Shah et al. [7], similar to the study by Eggmann et al. [38], does not present a systematic analysis of publications or topics. Instead, it focuses on the implications of large language models (*LLM*s) in medical applications. The authors emphasize that the development and use of *LLM*s in medicine should be actively managed. This involves ensuring the availability of relevant training data, clearly defining the expected benefits, and rigorously evaluating these benefits through real-world testing. This approach is crucial for ensuring that *LLM*s are integrated effectively into medical practice, maximizing their potential while addressing challenges related to their deployment.

Thirunavukarasu et al. [40] explored how applications such as ChatGPT are created and highlighted their use in clinical environments, but did not provide a systematic review of publications. The discussion includes the benefits and limitations of *LLM*s, emphasizing their potential to enhance clinical, educational, and research tasks in medicine. This review aims at providing an understanding for clinicians evaluating the integration of *LLM* technologies in healthcare to support patient and practitioner needs effectively. Finally, this study emphasizes medical and clinical aspects while overlooking technical and computer science perspectives.

The review by Blanco-González et al. [41] provides a comprehensive evaluation of the role of artificial intelligence (AI) in pharmaceutical research, highlighting both its potential benefits and the challenges it faces. Key discussions revolve around the use of AI for data augmentation, explainable AI, and the integration of AI with traditional experimental methods, with a strong focus on how these technologies can accelerate drug development. The review also explores the collaborative process between human authors and AI in scientific writing, specifically the use of ChatGPT, based on the GPT-3.5 language model, as an aid in manuscript drafting. The human authors' subsequent refinement and revision of the AI-generated content demonstrate the strengths and limitations of AI in academic writing. This highlights the necessity for careful human oversight to maintain the scientific integrity of AI-assisted contributions. The article concludes by providing insights into the evolving role of AI in enhancing scientific writing, underscoring the need for a balanced approach in leveraging AI's capabilities.

The paper by Xiao et al. [8] provides a comprehensive summary of transformer-based segmentation models in medical imaging, specifically applied to the abdominal organs, heart, brain, and lungs, drawing from studies published over the last two years. The authors analyzed the model architecture, focusing on the integration of transformers and their positioning within the segmentation model, as well as modifications made by researchers to enhance the model's performance. The study compared the segmentation results using the Dice evaluation metric, revealing that Unet-based transformer models are favored, with

transformers typically positioned in the encoder, as corroborated by 93 referenced works. The authors assert that their paper represents the first systematic review of medical image segmentation using transformer blocks.

The review by Harrer [42] focused on the ethical design of large language models, highlighting the risks of these tools being used to propagate misinformation and harmful content on a massive scale. The paper underscores the potential for *LLM*s to revolutionize healthcare and medical data workflows, while also addressing the operational principles, risks, and limitations associated with their use. The review advocates for a multidimensional framework, encompassing ethical, technical, and cultural considerations, to guide the responsible deployment of these technologies. It emphasizes the need for collaboration among all stakeholders—developers, regulators, and users—to ensure that *LLM*s are harnessed effectively for innovative and reliable applications in complex, evidence-driven fields such as healthcare.

The study by Sallam [6] is a systematic review conducted according to PRISMA guidelines to evaluate the utility and limitations of ChatGPT. Sallam analyzed English-language records from PubMed/MEDLINE and Google Scholar, including both published research and preprints, ultimately incorporating 60 studies. The review identifies significant improvements across various fields; however, it also highlights several concerns, such as ethical issues, copyright challenges, transparency, legal risks, and problems related to bias, plagiarism, hallucinated content, limited domain knowledge, incorrect citations, cybersecurity vulnerabilities, and the potential for infodemics. Notably, the review does not focus on technical aspects of ChatGPT's implementation.

The paper by Wang et al. [9] focused on the impact of GPT-4 on various healthcare applications, including radiology reporting, patient summaries, clinical decision support, and antimicrobial guidance. The study demonstrates that GPT-4 holds significant promise in enhancing these areas, but emphasizes the importance of adhering to ethical principles such as beneficence, transparency, and privacy. These principles are crucial to mitigate potential risks and ensure the responsible and equitable use of AI technologies in healthcare. The authors propose the development of a regulatory framework that involves collaboration between society, healthcare institutions, and AI developers to address the legal, humanistic, algorithmic, and information ethics concerns. The paper stresses that balancing the benefits and risks of GPT-4, as AI technology evolves rapidly, is essential for its effective and ethical integration into healthcare systems.

The study by Ray [43] offers a comprehensive overview of ChatGPT's origins, technological foundations, and its diverse applications in sectors like customer service, healthcare, and education. ChatGPT is portrayed as a transformative tool capable of reshaping traditional workflows by enhancing efficiency and improving accessibility in these domains. The review also critically addressed the challenges associated with its use, such as ethical concerns, algorithmic biases, data limitations, and safety issues. Ray underscored the necessity of implementing robust mitigation strategies to ensure the responsible and effective deployment of the technology. Future directions for ChatGPT were explored, particularly its integration with emerging technologies such as augmented reality (AR) and the Internet of Things (IoT), which could help bridge the digital divide. Ethical issues surrounding AI-assisted innovation were also discussed, emphasizing the importance of balancing AI's capabilities with human oversight and expertise. Despite ongoing ethical debates, ChatGPT's rapid adoption across academia, research, and industry underscores its transformative potential in the AI landscape.

Table 11. Summary of papers belonging to $Review_{New}$.

Title	Ref.	Source	Subject	AI Model	Date	Focus
The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective	[37]	Scopus	Business Management	CHATGpt	April. 2024	Applications
ChatGPT in healthcare: A taxonomy and systematic review	[10]	Scopus, PubMed	**Computer Science**, Medicine	CHATGpt	Mar 2024	Application, impact
Opportunities and challenges for ChatGPT and large language models in biomedicine and health	[11]	Scopus, PubMed	**Computer Science**, Medicine	CHATGpt	Jan. 24	Role, impact
Implications of large language models such as ChatGPT for dental medicine	[38]	Scopus, PubMed	Medicine (Dental)	CHATGpt	Oct. 2023	Non-technical Issues
Summary of ChatGPT-Related research and perspective towards the future of large language models	[39]	Scopus	Not present in SJR	CHATGpt	Sep. 2023	Non-technical Issues
Creation and Adoption of Large Language Models in Medicine	[7]	Scopus, PubMed	Not present in SJR	LLM	Sep. 2023	Clinical Issues
Large language models in medicine	[40]	Scopus, PubMed	Biochemistry, genetics	LLM	Jul. 2023	Clinical aspects
The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies	[41]	Scopus	Molecular Medicine, Drug Discovery, and Pharmaceutical Science	CHATGpt	Jun. 2023	Medical and clinical impact
Transformers in medical image segmentation: A review	[8]	Scopus	**Computer Science**, **Engineering** and Medicice	Transformer	Jul. 2023	AI model applied to medicine
Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine	[42]	Scopus, PubMed	Biochemistry Genetics and Molecular Biology and Medicine	LLM	Apr. 2023	Ethical and practical discussion
ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns	[6]	Scopus	health care systems, industry, technology, policy, and regulation	CHATGpt	Mar. 2023	benefit and limitation
Ethical Considerations of Using ChatGPT in Health Care	[9]	Scopus, PubMed	Health Informatics (Medicine)	CHATGpt	Aug. 2023	Impact of AI
ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope	[43]	Scopus	**Computer Science** and **Electrical and Electronic Engineering**	CHATGpt and AI	Apr. 2023	Technology and ethic
Shifting machine learning for healthcare from development to deployment and from models to data	[44]	Scopus, PubMed	Biochemistry, Genetics and Molecular Biology, Chemical Engineering (Bioengineering), **Computer Science Applications**, Biomedical **Engineering** and Medicine.	ML	Jul. 2022	Data management and AI technology

The study by Zhang et al. [44] examined the role of machine learning (ML) in healthcare, focusing on both its advancements and the challenges it faces. A central theme is the importance of data in developing and deploying ML models, which is crucial for automating physician tasks and enhancing clinical capabilities, including improved access to care. Topics discussed in the paper include the use of deep generative models and federated learning to augment datasets, leading to enhanced model performance. Additionally, the study highlighted the application of transformer models to manage larger datasets and enhance the analysis of clinical text, which is essential for tasks such as natural language processing (NLP) in healthcare. The review also delved into the challenges associated with the deployment of ML models in clinical settings, particularly the efficient delivery of data for real-time predictions. Zhang et al. also emphasized the issue of natural data shifts that can degrade model performance in dynamic clinical environments. Overall, this work provides a comprehensive review of how ML technologies are reshaping healthcare

through data-centric approaches while identifying key obstacles that need to be addressed for their successful integration into clinical practice.

The above analysis, given the nature of a citation count that is cumulative, is not able to capture very recent publications since they cannot accumulate enough citation to overcome the threshold. Although mechanisms have been implemented to mitigate this issue—we considered citation by year instead of the raw value—very recent publications, even those demonstrating a positive trend, cannot be included in the previous list. In order to make previous analysis more complete, we selected the papers published in 2024 that have more than 25 citation (that means more than 25 citation for year) and discussed those not already included in $Review_{New}$.

Islam et al. [45], with 34 citations at the time of writing, presented a comprehensive survey on transformers for deep learning tasks. The paper conducted a thorough analysis of highly effective models across five domains, proposed a taxonomy to classify these models according to their respective tasks, and explored future directions and challenges for transformer-based models. The survey covers publications from 2017 to 2022. The journal topic is computer science and applications.

Ref. [46], by Bhayana, cited 28 times, belongs to medical applications, specifically radiology, and it reviewed the limitations of LLMs and mitigation strategies, as well as potential uses of LLMs, including multimodal models.

Younis et al., in [47], published in a journal whose topics are biochemistry and biology, providing a systematic literature review categorizing 82 papers into eight major areas, including treatment and medicine, patient care, and medical imaging. However, it focused on the impact of chatbots from a clinical and medical point of view, highlighting the need for human judgment. This paper has been cited 27 times.

In Ref. [48], Chen and Esmaeilzadeh analyzed the current state of generative AI in healthcare, identifying opportunities and the privacy and security challenges posed by integrating these technologies into existing healthcare infrastructure. They proposed strategies for mitigating associated risks, emphasizing the importance of addressing security and privacy threats to ensure the safe and effective use of generative AI systems in healthcare. The journal subject is medicine and the citation count is 25.

Schukow et al. in [49] explored the current status and understanding of ChatGPT's potential applications in routine diagnostic pathology. They emphasized that any use of the ChatGPT knowledge, coming from a large but not verified sources, at the patient care level must be carefully merged with established medical information sources and expertise. The authors suggested that, with the ever-expanding knowledge base required for personalized, precision anatomic pathology, improved technologies like future versions of ChatGPT could serve as key allies to diagnosticians. The citation count is 25 and the journal subject is strongly connected with medicine and clinical applications.

In summary, the majority of reviews have primarily addressed ethical considerations regarding the application of AI in medicine rather than delving into technical aspects, algorithms, or model architectures. These reviews have emphasized clinical implications and the societal responsibilities tied to AI's integration into healthcare. Moreover, many of the papers are published in journals specifically dedicated to medicine or closely related fields. Notably, some of these journals are not currently indexed or classified in databases like SJR, potentially reflecting their niche focus or emerging status in academic publishing.

6. Discussions and Conclusions

As shown in the literature, numerous examples of the applicability of *LLM*s in the medical field have been presented. However, the long-term success of these technologies faces significant barriers, particularly in terms of data privacy and ethical challenges.

Ensuring compliance with stringent data protection regulations such as GDPR and HIPAA remains a major concern, as healthcare data are highly sensitive. Furthermore, ethical issues, including potential algorithmic bias and the need for transparent decision making by AI, present considerable obstacles. Addressing these challenges will be crucial for the future scalability and reliability of *LLM*s in the healthcare context. To overcome these hurdles, it is essential that healthcare professionals and policymakers adopt precautionary approaches by developing clear guidelines for the ethical and regulatory use of *LLM*s.

In particular, healthcare professionals should be trained to understand the limitations and potential of language models, ensuring that the use of these tools does not compromise the quality of clinical decisions. Policymakers, on the other hand, should establish regulations that protect patient privacy while simultaneously promoting technological innovation responsibly. Ethical considerations, such as managing algorithmic bias and ensuring transparency in automated decision-making processes, must be an integral part of healthcare policies.

The results reveal a steadily growing trend in the adoption of *LLM*s, with three distinct periods identified, the most recent of which shows an exponential increase in publications. A significant proportion of the studies focuses on practical applications of *LLM*s in medicine, highlighting their increasing relevance and utility in real-world scenarios.

Collaboration networks demonstrate the presence of central hubs, often located in major research institutions of global importance. However, notable contributions are also observed from smaller research centers, reflecting a diverse and distributed landscape of scientific efforts. At a national level, collaboration hubs are primarily concentrated in the United States and other English-speaking countries, although smaller nations also play a key role in specific collaborations, showcasing their impactful contributions to the field.

Future work will extend the analysis of collaborations to identify research communities and investigate their significance, with the aim of uncovering underlying structures and dynamics that drive scientific advancements in the application of *LLM*s to medicine.

Regarding future developments, the analysis of the literature could evolve into a more collaborative and dynamic approach, where researchers, professionals, and policymakers work together to explore the impact of *LLM* usage in medicine. A deeper analysis of collaborations between various stakeholders over time could highlight significant changes in research dynamics, revealing how international alliances, technological innovations, and healthcare policies have influenced the evolution of *LLM* technologies. Additionally, such an analysis could uncover how socio-political transformations have contributed to the adoption and direction of research on the use of *LLM*s in healthcare. Growing awareness of privacy, ethical, and security issues, as well as interest in advanced technological solutions, may manifest differently at a global level, leading to political and social implications that could shape the evolution of digital medicine. Analyzing publications in the healthcare field could reveal how regulatory approaches and public policies have responded to these developments, offering new insights to guide future research.

By overcoming these challenges, the adoption of *LLM*s could profoundly transform healthcare systems, improving not only operational efficiency but also the quality and equity of treatments, while enabling a more adequate response to socio-political changes that influence both medicine and technology. This study provides a quantitative review of the scientific literature on the application of large language models in the medical field, offering an in-depth analysis of emerging trends, research collaborations, and thematic focuses.

Author Contributions: Methodology, V.C. and M.M.; Software, V.C. and M.M.; Validation, V.C. and M.M.; Investigation, M.M.; Data curation, V.C. and M.M.; Writing – original draft, V.C. and M.M.; Writing – review & editing, V.C. and M.M.; Funding acquisition, V.C.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the UDMA project, CUP: G69J18001040007.

Data Availability Statement: The original contributions presented in this 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.

References

1. Hager, P.; Jungmann, F.; Holl, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* **2024**, *30*, 2613–2622. [CrossRef]
2. Goyal, S.; Rastogi, E.; Rajagopal, S.P.; Yuan, D.; Zhao, F.; Chintagunta, J.; Naik, G.; Ward, J. Healai: A healthcare llm for effective medical documentation. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining, Merida, Mexico, 4–8 March 2024; pp. 1167–1168.
3. Chen, S.; Guevara, M.; Moningi, S.; Hoebers, F.; Elhalawani, H.; Kann, B.H.; Chipidza, F.E.; Leeman, J.; Aerts, H.J.; Miller, T.; et al. The effect of using a large language model to respond to patient messages. *Lancet Digit. Health* **2024**, *6*, e379–e381. [CrossRef] [PubMed]
4. Carchiolo, V.; Malgeri, M.; Sapari, L.S. Conversational Agent for Handling Health Report Inquiries. In Proceedings of the 16th International Conference on Management of Digital Ecosystems (MEDES), Naples, Italy, 18–20 November 2024; Springer: Berlin/Heidelberg, Germany; 2024; Communications in Computer and Information Science.
5. Benary, M.; Wang, X.D.; Schmidt, M.; Soll, D.; Hilfenhaus, G.; Nassir, M.; Sigler, C.; Knödler, M.; Keller, U.; Beule, D.; et al. Leveraging large language models for decision support in personalized oncology. *JAMA Netw. Open* **2023**, *6*, e2343689. [CrossRef] [PubMed]
6. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef]
7. Shah, N.H.; Entwistle, D.; Pfeffer, M.A. Creation and Adoption of Large Language Models in Medicine. *JAMA* **2023**, *330*, 866–869. [CrossRef] [PubMed]
8. Xiao, H.; Li, L.; Liu, Q.; Zhu, X.; Zhang, Q. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control.* **2023**, *84*, 104791. [CrossRef]
9. Wang, C.; Liu, S.; Yang, H.; Guo, J.; Wu, Y.; Liu, J. Ethical Considerations of Using ChatGPT in Health Care. *J. Med. Internet Res.* **2023**, *25*, e48009. [CrossRef]
10. Li, J.; Dada, A.; Puladi, B.; Kleesiek, J.; Egger, J. ChatGPT in healthcare: A taxonomy and systematic review. *Comput. Methods Programs Biomed.* **2024**, *245*, 108013. [CrossRef]
11. Tian, S.; Jin, Q.; Yeganova, L.; Lai, P.T.; Zhu, Q.; Chen, X.; Yang, Y.; Chen, Q.; Kim, W.; Comeau, D.C.; et al. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings Bioinform.* **2024**, *25*, bbad493. [CrossRef]
12. Carchiolo, V.; Malgeri, M. Navigating the AI Timeline: From 1995 to Today. In Proceedings of the 13th International Conference on Data Science, Technology and Applications, DATA 2024, Dijon, France, 9–11 July 2024; pp. 577–584. [CrossRef]
13. Elsevier Developer Portal. Elsevier Developer—API Service Agreement. Available online: https://dev.elsevier.com/academic_research_scopus.html (accessed on 17 December 2024).
14. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
15. Dernbach, S.; Agarwal, K.; Zuniga, A.; Henry, M.; Choudhury, S. GLaM: Fine-Tuning Large Language Models for Domain Knowledge Graph Alignment via Neighborhood Partitioning and Generative Subgraph Encoding. In Proceedings of the 2024 AAAI Spring Symposium Series, Stanford, CA, USA, 25–27 March 2024; The AAAI Press: Washington, DC, USA, 2024; Volume 3, pp. 82–89.
16. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
17. Dale, R. GPT-3: What is it good for? *Nat. Lang. Eng.* **2021**, *27*, 113–118. [CrossRef]
18. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
19. Driess, D.; Xia, F.; Sajjadi, M.S.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. Palm-e: An embodied multimodal language model. *arXiv* **2023**, arXiv:2303.03378.
20. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Hanna, E.B.; Bressand, F.; et al. Mixtral of experts. *arXiv* **2024**, arXiv:2401.04088.

21. Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv* **2022**, arXiv:2204.05862.
22. Powles, J.; Hodson, H. Google DeepMind and healthcare in an age of algorithms. *Health Technol.* **2017**, *7*, 351–367. [CrossRef]
23. 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. *Briefings Bioinform.* **2022**, *23*, bbac409. [CrossRef] [PubMed]
24. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Flores, M.G.; Zhang, Y.; et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv* **2022**, arXiv:2203.03540.
25. 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**. [CrossRef] [PubMed]
26. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *Acm Trans. Comput. Healthc. (HEALTH)* **2021**, *3*, 1–23. [CrossRef]
27. Taylor, R.; Kardas, M.; Cucurull, G.; Scialom, T.; Hartshorn, A.; Saravia, E.; Poulton, A.; Kerkez, V.; Stojnic, R. Galactica: A large language model for science. *arXiv* **2022**, arXiv:2211.09085.
28. Alsentzer, E.; Murphy, J.R.; Boag, W.; Weng, W.H.; Jin, D.; Naumann, T.; McDermott, M. Publicly available clinical BERT embeddings. *arXiv* **2019**, arXiv:1904.03323.
29. 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] [PubMed]
30. Church, K.W. Word2Vec. *Nat. Lang. Eng.* **2017**, *23*, 155–162. [CrossRef]
31. Wu, S.; Koo, M.; Blum, L.; Black, A.; Kao, L.; Scalzo, F.; Kurtz, I. A comparative study of open-source large language models, gpt-4 and claude 2: Multiple-choice test taking in nephrology. *arXiv* **2023**, arXiv:2308.04709.
32. Saeidnia, H.R. Welcome to the Gemini era: Google DeepMind and the information industry. *Library Hi Tech News* **2023** [CrossRef]
33. Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; et al. Bloom: A 176b-parameter open-access multilingual language model *arXiv* **2023**, arXiv:2211.05100. [CrossRef]
34. Khalid, H.; Hussain, M.; Al Ghamdi, M.A.; Khalid, T.; Khalid, K.; Khan, M.A.; Fatima, K.; Masood, K.; Almotiri, S.H.; Farooq, M.S.; et al. A comparative systematic literature review on knee bone reports from mri, x-rays and ct scans using deep learning and machine learning methodologies. *Diagnostics* **2020**, *10*, 518. [CrossRef] [PubMed]
35. Carchiolo, V.; Grassia, M.; Malgeri, M.; Mangioni, G. Co-authorship Networks Analysis to Discover Collaboration Patterns among Italian Researcher. *Future Internet* **2022**, *14*. [CrossRef]
36. Hardeniya, N.; Perkins, J.; Chopra, D.; Joshi, N.; Mathur, I. *Natural Language Processing: Python and NLTK*; Packt Publishing Ltd.: Birmingham, UK, 2016.
37. Kanbach, D.K.; Heiduk, L.; Blueher, G.; Schreiter, M.; Lahmann, A. The GenAI is out of the bottle: Generative artificial intelligence from a business model innovation perspective. *Rev. Manag. Sci.* **2024**, *18*, 1189–1220. [CrossRef]
38. Eggmann, F.; Weiger, R.; Zitzmann, N.U.; Blatz, M.B. Implications of large language models such as ChatGPT for dental medicine. *J. Esthet. Restor. Dent.* **2023**, *35*, 1098–1102. [CrossRef]
39. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology* **2023**, *1*, 100017. [CrossRef]
40. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef] [PubMed]
41. Blanco-González, A.; Cabezón, A.; Seco-González, A.; Conde-Torres, D.; Antelo-Riveiro, P.; Piñeiro, A.; Garcia-Fandino, R. The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. *Pharmaceuticals* **2023**, *16*, 891. [CrossRef]
42. Harrer, S. Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *eBioMedicine* **2023**, *90*, 104512. [CrossRef]
43. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber Phys. Syst.* **2023**, *3*, 121–154. [CrossRef]
44. Zhang, A.; Xing, L.; Zou, J.; Wu, J.C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng.* **2022**, *6*, 1330–1345. [CrossRef]
45. Islam, S.; Elmekki, H.; Elsebai, A.; Bentahar, J.; Drawel, N.; Rjoub, G.; Pedrycz, W. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Syst. Appl.* **2024**, *241*, 122666. [CrossRef]
46. Bhayana, R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology* **2024**, *310*, e232756. [CrossRef]

47. Younis, H.A.; Eisa, T.A.E.; Nasser, M.; Sahib, T.M.; Noor, A.A.; Alyasiri, O.M.; Salisu, S.; Hayder, I.M.; Younis, H.A. A Systematic Review and Meta-Analysis of Artificial Intelligence Tools in Medicine and Healthcare: Applications, Considerations, Limitations, Motivation and Challenges. *Diagnostics* **2024**, *14*, 109. [CrossRef]
48. Chen, Y.; Esmaeilzadeh, P. Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. *J. Med. Internet Res.* **2024**, *26*, e53008. [CrossRef]
49. Schukow, C.; Smith, S.; Landgrebe, E.; Parasuraman, S.; Folaranmi, O.; Paner, G.; Amin, M. Application of ChatGPT in Routine Diagnostic Pathology: Promises, Pitfalls, and Potential Future Directions. *Adv. Anat. Pathol.* **2024**, *31*, 15–21. [CrossRef] [PubMed]