Engineering Reports

WILEY

Engineering Reports Open Access

REVIEW OPEN ACCESS

Systematic Literature Review of LLM-Large Language Model in Medical: Digital Health, Technology and Applications

Imrus Salehin[1,2] | Md Tomal Ahmed Sajib[1] | Nazmul Huda Badhon[1] | Md Sakibul Hassan Rifat[3] | Nazrul Amin[4] | Nazmun Nessa Moon[1]

[1]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh | [2]Department of Information Convergence Engineering, Pusan National University, Busan, Republic of Korea | [3]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh | [4]Department of Computer Engineering, Jeju National University, Jeju-si, Republic of Korea

Correspondence: Nazmul Huda Badhon (badhon15-4772@diu.edu.bd)

Received: 26 April 2025 | **Revised:** 27 June 2025 | **Accepted:** 18 August 2025

Funding: The authors received no specific funding for this work.

Keywords: AI | Health | LLM | Medical AI | NLP

ABSTRACT

Large language models (LLMs), like the GPT series, have recently emerged as transformative tools in the medical field due to their human-like language generation and understanding. This systematic review examines the evolution, applications, and challenges of medical LLMs in digital health and clinical technology. A structured search was conducted across ScienceDirect, PubMed, Scopus, and manual sources from 2007 to 2025, following PRISMA 2020 guidelines. After applying inclusion and exclusion criteria, 185 studies were selected from an initial pool of 698 papers. Among these, 30 representative studies were analyzed in-depth based on their relevance, methodological quality, and contribution to diverse LLM applications in health care. Most research centered on GPT-based models, with over 81% demonstrating strong performance in language generation, diagnostic assistance, and clinical documentation, based on automated metrics and human feedback. Notably, some models achieved up to 90% satisfaction from healthcare professionals. The findings reveal LLM's potential to enhance patient interaction, decision support, and overall healthcare efficiency. This review contributes by synthesizing key advancements, assessing model performance, and outlining ethical challenges such as trust, privacy, and safe deployment. It offers novel insights for researchers and practitioners seeking to adopt or improve LLM integration in health care. Future directions include improving transparency, developing domain-specific models, and establishing regulatory frameworks for responsible use.

1 | Introduction

Recent advancements in large language models (LLMs) [1, 2] like PaLM, LLaMA [3], the GPT series [4], and ChatGLM [5] have significantly pushed the boundaries of performance across various natural language processing (NLP) tasks, including text generation, summarization, and question answering. Building on the success of general LLMs, various initiatives have emerged to adapt these models specifically for the medical domain, leading to the creation of specialized medical LLMs. For instance, MedPaLM-2 [6] and MedPrompt [7], built on PaLM and GPT-4 respectively, have demonstrated impressive performance on the United States Medical Licensing Examination (USMLE) [8], achieving competitive accuracies of 86.5 and 90.2, compared to human expert's 87.013. Medical LLMs are increasingly being leveraged to improve diagnostics, enhance patient interactions, and optimize clinical workflows. Medical LLMs excel at summarizing clinical notes accurately and efficiently extracting essential information from complex medical documents. This allows healthcare professionals to quickly review and understand patient data, streamlining the decision-making process. In addition, medical LLMs play a crucial role in administrative tasks like clinical documentation, medical research, and patient care. Medical LLMs efficiently analyze large volumes of medical literature, clinical records, and scientific publications. This capability allows them to provide valuable insights, support accurate diagnoses, and assist in treatment planning. As a result, healthcare professionals are better equipped to make informed decisions, ultimately improving patient care.

In recent years, the emergence of Web 3.0 [9] and the Internet of Behavior [10] has paved the way for tremendous advances in artificial intelligence and data science [11, 12], as noted in the emergence of Large Language Models (LLMs). These models, like GPT and BERT, represent a remarkable leap in AI technology. LLMs are specifically designed to work on natural language writing and are developed based on an in-depth learning program. Language structures are learned, and proficiency is gained through prior training on a broad range of textual data, resulting in an excellent performance in a wide range of Natural Language Processing (NLP) tasks, including text composition, synthesis, interpretation, and question. The advancement of LLMs is due to two key factors. One significant factor is the availability of vast pre-training datasets. These large data sets enable LLMs to capture a wide range of linguistic knowledge and patterns by drawing from richly annotated sources across the Internet. Second, advances in computer hardware and technology have been exponential. Increases in computing power, the development of specialized hardware, such as GPUs and TPUs, and advances in parallel processing techniques have provided important tools for training and efficiently implementing large language models.

Despite the promising progress and comprehensive reviews, the potential future directions of medical LLMs remain underexplored, particularly in aligning their development with the complex demands of clinical settings, an essential step for improving patient care and advancing medical research. LLMs are increasingly being applied in areas such as diagnostic support, patient communication, medical research synthesis, and documentation, with the potential to improve outcomes and reduce clinical workloads. However, integration remains challenging due to issues of accuracy, ethical compliance, data privacy, and adaptability to evolving medical knowledge. Compared to recent studies on medical LLMs, a more focused literature review is necessary to explore both technical (e.g., medical tasks, data, evaluation, and algorithms) and social impacts (e.g., application, trustworthiness, and safety). Moreover, reliance on large-scale but potentially biased data sets introduces fairness and transparency concerns that must be addressed. Due to the rapid growth of medical LLMs, it is challenging for the research community to carry out a comprehensive and detailed examination of current models, covering their historical background and technological innovations. Aligning medical LLMs with clinician values poses difficulties, particularly due to the risk of generating harmful, false, or toxic content. Trustworthy and ethical guidelines must be considered to ensure fairness, accountability, privacy, and robustness. A structured synthesis of current trends, capabilities, and limitations is essential to guide the responsible development of LLMs that meet healthcare demands and ethical standards.

Furthermore, the integration of LLMs into different roles has transformed many industries. For example, medical LLMs are being used in the medical field to improve diagnosis, facilitate patient communication, and improve clinical performance. Many medical journals, clinical articles, and scientific papers can be searched. This paper explores the impact of Large Language Models (LLMs) in health care, focusing on their diverse applications and potential. Section 2 outlines the systematic literature review approach, including the article search and selection process. Section 3 covers key LLM models like GPT, BERT, and others, alongside a modified table comparing these models. Section 4 categorizes LLM applications, including clinical decision support, medical text processing, and patient interaction. Section 5 delves into specific uses, such as diagnostic assistance, predictive modeling, drug discovery, and disease detection. Section 6 discusses the challenges of LLM implementation, like data privacy concerns, model interpretability, and deployment costs. Section 7 discusses key findings, while Section 8 concludes by highlighting the transformative potential of LLMs in improving healthcare outcomes and paving the way for future innovations.

2 | Systematic Literature Review Approach

A systematic literature review is a structured and methodical approach aimed at comprehensively analyzing, synthesizing, and critically appraising existing research within a specific domain. In the context of Large Language Models in medical and digital health technology, we follow a predefined organized process of identifying relevant studies, assessing the quality of evidence, and synthesizing findings to address key research questions. The methodology we follow adheres to rigorous protocols, such as PRISMA [13], which play a critical role in ensuring the transparency, consistency, and replicability of the systematic review process (Figure 1). PRISMA provides a standardized framework for defining the review's objectives, conducting an exhaustive literature search, and transparently documenting the inclusion and exclusion of studies (Table 1). This structured approach minimizes bias and enhances the reliability of the review, enabling other researchers to reproduce the process. This section outlines the methodology employed, including the article search and selection process, criteria for inclusion and exclusion, and the overall literature selection workflow.

2.1 | Step 1: Article Search and Selection Process

The initial phase of the systematic literature review, often referred to as the data processing phase, involves identifying and selecting relevant studies that align with the research objectives. This critical step ensures that the review remains comprehensive, capturing the most pertinent literature on LLM in medical and digital health. To achieve an exhaustive search, the aforementioned databases and platforms are utilized:

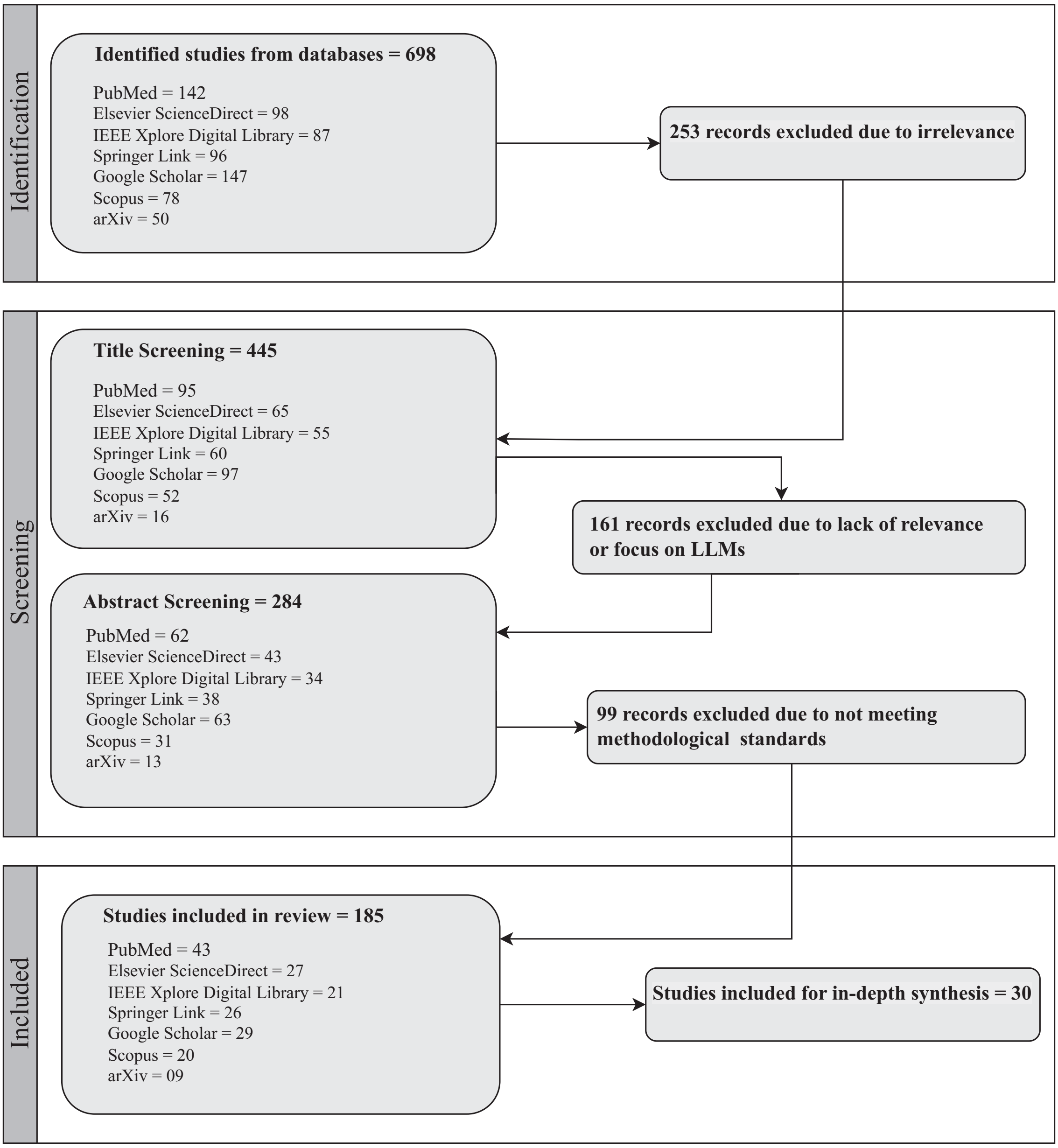

FIGURE 1 | PRISMA-based screening flowchart illustrating the systematic review process. The diagram outlines the number of studies identified, screened, and included at each stage, along with reasons for exclusion. A total of 698 records were retrieved from multiple databases, with 185 studies ultimately included in the review and 30 selected for in-depth synthesis.

TABLE 1 | Number of literature selections by database.

Data sources	Keyword search	Title screening	Abstract screening	Full-text screening
PubMed	142	95	62	43
Elsevier ScienceDirect	98	65	43	27
IEEE Xplore Digital Library	87	55	34	21
Springer Link	96	60	38	26
Google Scholar	147	97	63	39
Scopus	78	52	31	20
arXiv	50	16	13	9
Total	698	445	284	185

1. PubMed (https://pubmed.ncbi.nlm.nih.gov)
2. Elsevier ScienceDirect (https://www.sciencedirect.com)
3. IEEE Xplore Digital Library (https://ieeexplore.ieee.org)
4. Springer Link (https://link.springer.com)
5. Google Scholar (https://scholar.google.com)
6. Scopus (https://www.scopus.com)
7. arXiv (https://arxiv.org)

To ensure comprehensive coverage of the literature, recognizing the importance of capturing studies from diverse disciplines and perspectives, multiple databases are employed. The selection of databases is guided by the research question and subject area, focusing on reputable platforms to enhance the quality of sources.

In conducting the systematic literature review, a well-defined search strategy is implemented, incorporating specific keywords and Boolean operators to optimize the retrieval of relevant articles.

- Large Language Models: Large Language Model, LLM, Generative Pre-trained Transformer, GPT, BERT, XLNet, RoBERTa, ALBERT, Megatron
- Medical Applications: Medical, Healthcare, Clinical Decision Support, Medical Text Processing, Drug Discovery, Pharmacology, Epidemiology
- Digital Health Technology: Digital Health, Telemedicine, Health Informatics, Patient Interaction, Health Chatbots
- Integration Terms: NLP, Natural Language Processing, Machine Learning, Deep Learning, Artificial Intelligence

We utilize these keywords to ensure that we effectively capture the fundamental concepts surrounding the applications, advancements, and challenges of Large Language Models in the medical and digital health domains. A broad spectrum of research papers aligned with these key areas is examined, contributing to an in-depth understanding of LLMs in the context of medical and digital health technologies.

2.2 | Step 2: Criteria for Inclusion and Exclusion

In the process of selecting 185 papers from a data set of 698 scientific articles, specific requisites were established to ensure the selection of the most relevant and high-quality studies. These requisites consist of well-defined criteria that guide the decision-making process for determining whether a paper should be included or excluded from the review. Below are the detailed criteria for inclusion and exclusion.

Requisites for Inclusion:

- *Relevance to Core Concepts*: The review paper directly addresses one or more aspects of Large Language Models in the context of medical or digital health technology.
- *Comprehensive Coverage*: The review paper provides a thorough analysis or synthesis of the literature on LLM applications, methodologies, or impacts in the medical and digital health fields.
- *Methodological Rigor*: The review paper demonstrates a sound methodological approach, such as empirical research, systematic reviews, or meta-analyses.
- *Recent Publication*: Preference papers published within a specific timeframe to ensure the inclusion of the most up-to-date research on LLM, medical, and digital health.
- *Language*: Only studies published in English are considered to maintain consistency and comprehensibility.

Requisites for Exclusion:

- *Irrelevance to Core Concepts*: Review papers not about LLM or their applications in medical and digital health technology are excluded.
- *Limited Scope or Depth*: Papers offering superficial insights or focusing on unrelated aspects are excluded in favor of more comprehensive studies.
- *Methodological Flaws*: Papers with significant methodological weaknesses, such as small sample sizes, lack of control groups, or poor data analysis, are excluded to ensure reliability.
- *Poor Quality or Inadequate Reporting*: Papers with incomplete data, unclear methodologies, or low-quality content related to LLM in the target domains are excluded to maintain the overall quality of the review.

By applying these inclusion and exclusion criteria based on the fundamental concepts of LLMs, medical technology, digital health, and NLP, the selection process ensures that the 185 chosen papers offer relevant, comprehensive, and high-quality insights into these specific fields. This rigorous selection supports further analysis and synthesis within the research domain.

2.3 | Step 3: Literature Selection Process

The literature selection process involves multiple stages of screening to systematically narrow down the pool of identified studies to those most relevant and robust. The process comprises four main stages: keyword search, title screening, abstract screening, and full-text screening.

- *Keyword Search*: An initial search using the defined keywords across all selected databases yielded a total of 698 papers potentially relevant to the review.
- *Title Screening*: The titles of the 698 papers were meticulously reviewed to assess their relevance to LLMs in medical and digital health. This stage resulted in the exclusion of 253 papers that were deemed irrelevant, leaving 445 papers for further screening.
- *Abstract Screening*: The abstracts of the remaining 445 papers were examined to evaluate their alignment with the

research objectives. Based on this screening, 161 papers were excluded due to lack of relevance or insufficient focus on LLMs, resulting in 284 papers proceeding to the full-text screening stage.

- *Full-Text Screening*: The full texts of the 284 papers were thoroughly assessed against the inclusion and exclusion criteria. This comprehensive evaluation led to the exclusion of 99 papers that did not meet the methodological or content standards, culminating in a final selection of 185 high-quality studies for inclusion in the systematic literature review.

The final set of 185 papers encompasses a diverse range of studies, including empirical research, systematic reviews, and meta-analyses, providing a robust foundation for synthesizing the current state of knowledge on Large Language Models in medical and digital health technology. Of the 185 papers included after full-text screening, a subset of 30 studies was selected for detailed synthesis and analysis. These 30 articles were chosen based on a combination of factors, including (i) their methodological rigor, (ii) coverage of different application domains (e.g., diagnostics, summarization, patient interaction), and (iii) impact or novelty of the citation.

This selective deep-dive allowed us to capture meaningful patterns and insights while keeping the review focused and coherent. The remaining studies were referenced to support general trends and provide context throughout the review.

3 | Related Concept of Large Language Model

Key concepts and architectures in Large Language Models are foundational to advancements in Natural Language Processing. Prominent LLMs include GPT (Generative Pre-trained Transformer), which utilizes autoregressive generation to predict and generate text based on prior tokens, making it effective for tasks like text completion and conversation. BERT (Bidirectional Encoder Representations from Transformers) uses masked language modeling to capture deep bidirectional context, facilitating performance improvements in understanding sentence structure and meaning. XLNet extends BERT and GPT principles by incorporating permutation-based training, enabling it to capture dependencies across word order for enhanced language understanding. RoBERTa is an optimized variant of BERT, fine-tuned through extensive training on larger datasets, achieving superior results on NLP benchmarks. ALBERT further enhances efficiency by implementing parameter-sharing techniques to reduce model size without sacrificing performance, making it suitable for resource-limited environments. Finally, Megatron is a large-scale transformer model optimized for distributed training across multiple GPUs, enabling it to scale up to billions of parameters for high-performance NLP applications.

Table 2 provides a statistical summary of LLMs with more than 10B parameters, covering aspects such as release dates, model sizes, base architectures, pre-training data scale, and hardware requirements. Only models with publicly accessible technical papers are listed, with classifications as "Publicly Available" if model checkpoints are accessible, or "Closed Source" otherwise. This comparative table offers readers a high-level overview of key differences across leading LLMs and highlights which models may be more suitable for adaptation in medical and digital health applications. Across these models, a clear trend emerges toward increasing parameter size and pretraining data scale, particularly in proprietary models like GPT-4 and Gemini, which often remain closed-source. While such models demonstrate cutting-edge capabilities, their closed nature and high computational demands pose limitations for clinical adoption. By contrast, publicly available models (e.g., BERT variants) offer more flexibility for fine-tuning and integration in resource-constrained healthcare settings. This trade-off between performance and accessibility is central to the ongoing evolution of medical LLM and is further discussed in the Challenges section.

This section covers broader Generative Models, which are designed to create new content by learning patterns from extensive data sets. These models generate coherent and contextually relevant text, images, or other media by sampling from learned distributions. Autoregressive models are also discussed, focusing on their sequential generation approach, where each token is predicted based on preceding ones, a feature commonly seen in models like GPT. Lastly, Natural Language Processing (NLP) is examined as the overarching domain, contextualizing LLMs in applications such as machine translation, sentiment analysis, question-answering, and conversational AI, showcasing their potential to transform diverse language-based tasks.

3.1 | Large Language Model (LLM)

Large Language Models (LLMs) are state-of-the-art AI systems designed to comprehend and generate human-like text in different applications [14]. Each has architectural and functional elements that define its work and design, and these elements are based on several components that allow for effective text processing [15]. Most LLMs are developed based on the transformer architecture, which transformed the field of natural language processing (NLP) because the models can efficiently process large amounts of text [16]. The heart of LLMs is called the attention mechanism, which enables the model to determine how much attention it should pay to one token versus another [17]. Self-attention operates over individual sequences while cross-attention is found in both encoder-decoder architectures [18]. To enhance the processing of extended sequences, sparse attention, and flash attention enhance the computational effectiveness. Furthermore, non-linearity is introduced by activation functions like ReLU, GeLU, and GLU variations [19], which improve the model's capacity to recognize and understand intricate, non-linear patterns in the data [20]. Layer normalization is critical to the training process of LLMs, more so in deep transformer architectures, concerned with approximately parallelizing LLMs on data, tensors, and models, where these models are trained on large data sets via utilizing distributed training [21]. LLM architectures from encoder-decoder [22] to causal decoders and a mixture of experts make it versatile for language generation and translation tasks. These models are pre-trained for generalization [23] through masked and prefix language modeling and, in terms of scaling laws, larger models are seen to yield improved performance [24]. Figure 2 depicts the basic architecture of an LLM using a transformer.

TABLE 2 | A statistical summary of large language models.

Access	Model	Release date	Base model	Size (B)	Pre-train data scale	Latest data time	Training time	Hardware
Closed	PaLM2	May, 2023	—	16	100B Tokens	—	—	—
	GPT-4	March, 2023	—	—	—	—	—	—
	Flan-U-PaLM	October, 2022	U-PaLM	540	—	—	—	—
	Flan-PaLM	October, 2022	PaLM	540	—	—	37 h	512 TPU v4
	Flan-PaLM	October, 2022	PaLM	540	—	—	5 Days	512 TPU v4
	WeLM	September, 2022	—	10	300B Tokens	—	24 Days	128 A100 40G
	Sparrow	September, 2022	—	70	—	—	—	64 TPU v3
	AlexaTM	August, 2022	—	20	1.3T Tokens	—	120 Days	128 A100
	PaLM	April, 2022	—	540	780B Tokens	—	—	6144 TPU v4
	InstructGPT	March, 2022	GPT-3	175	—	—	—	—
	AlphaCode	February, 2022	—	41	967B Tokens	July, 2021	—	—
	MT-NLG	January, 2022	—	530	270B Tokens	—	—	4480 80G A100
	LaMDA	January, 2022	—	137	768B Tokens	—	58 Days	1024 TPU v3
	GLaM	January, 2022	—	1200	280B Tokens	—	24 Days	1024 TPU v4
	Gopher	December, 2021	—	260	300B Tokens	—	38.3 Days	4096 TPU v3
	WebGPT	December, 2021	GPT-3	175	—	—	—	—
	FLAN	September, 2021	LaMDA-PT	137	—	—	60 h	128 TPU v3
	HyperCLOVA	September, 2021	—	82	300B Tokens	—	13.4 Days	1024 A100
	ERNIE 3.0	July, 2021	—	10	375B Tokens	—	—	384 V100
	Codex	July, 2021	GPT-3	12	100B Tokens	May, 2020	—	—
	GShared	June, 2020	—	600	1T Tokens	—	4 Days	2048 TPU v3
	GPT-3	May, 2020	—	175	300B Tokens	—	—	
Open	Skywork	October, 2023	—	13	3.2T Tokens	—	—	512 80G A800
	FLM	September, 2023	—	101	311B Tokens	—	22 Days	192 A800
	LLaMA2	July, 2023	—	70	2T Tokens	—	—	2000 80G A100
	StarCoder	May, 2023	—	15.5	1T Tokens	—	—	512 40G A100
	CodeGen2	May, 2023	—	16	400B Tokens	—	—	—
	Pythia	April, 2023	—	12	300B Tokens	—	—	256 40G A100
	LLaMA	February, 2023	—	65	1.4T Tokens	—	21 Days	2048 80G A100
	OPT-IML	December, 2022	OPT	175	—	—	—	128 40G A100
	BLOOMZ	November. 2022	BLOOM	176	—	—	—	128 40G A100
	mT0	November, 2022	mT5	13	—	—	—	—
	BLOOM	November, 2022	—	176	366B Tokens	—	105 Days	384 80G A100
	Flan-T5	October, 2022	T5	—	—	—	—	
	GLM	October, 2022	—	130	400B Tokens	—	60 Days	786 40G A100
	CodeGeeX	September. 2022	—	13	850B Tokens	—	60 Days	1536 Ascand 910
	UL2	May, 2022	—	20	1T Tokens	April, 2019	—	512 TPU v4
	Tk-Instruct	April, 2022	T5	11	—	—	4 h	256 TPU v3
	GPT-NeoX	April, 2022	—	20	825GB	—	—	96 40G A100
	CodeGEn	March, 2022	—	16	577B Tokens	—	—	—
	T0	October, 2021	T5	11	—	—	27 h	512 TPU v3
	CPM-2	June, 2021	—	198	2.6TB	—	—	—
	mT5	October, 2020	—	13	1T Tokens	—	—	—
	T5	October, 2019	—	11	1T Tokens	April, 2019	—	1024 TPU v3

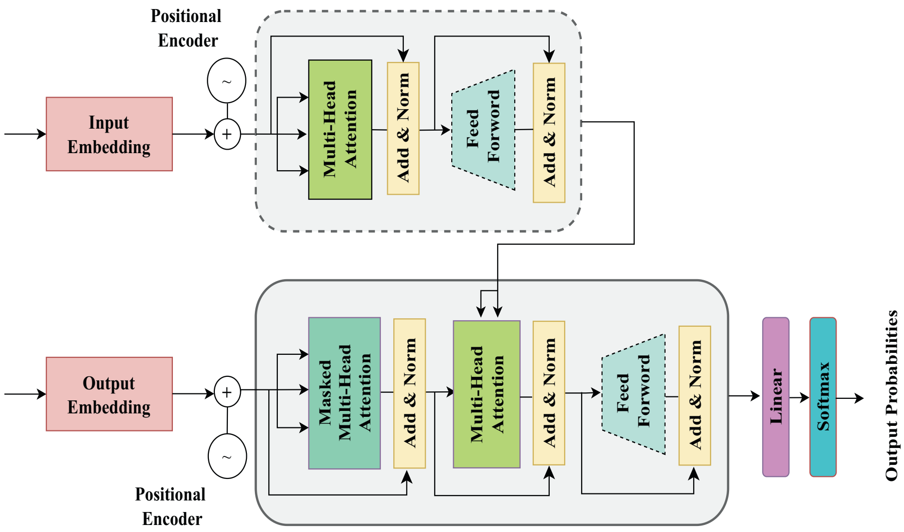

FIGURE 2 | Overview of the fundamental transformer architecture that underpins large language models (LLMs). This figure illustrates how attention mechanisms and encoder-decoder layers interact to process sequential input dataŮessential for tasks such as language generation, translation, and medical text comprehension.

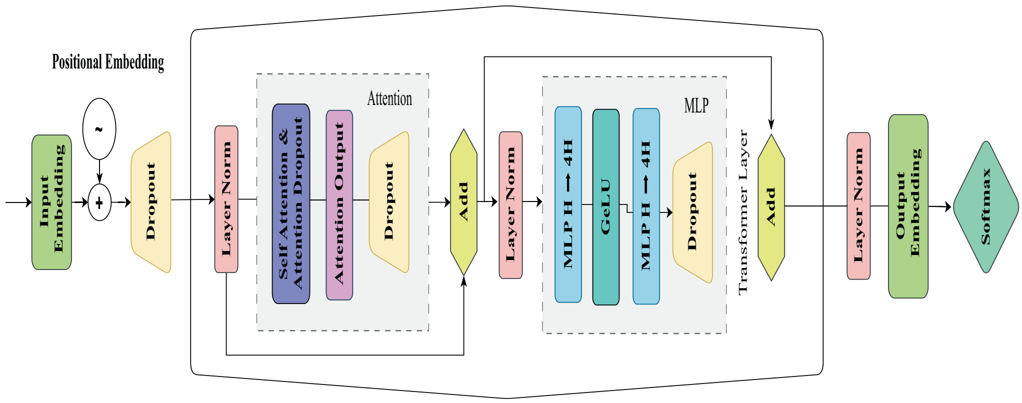

FIGURE 3 | Schematic representation of the GPT architecture, showing the stacked transformer decoder layers used for autoregressive language modeling. This architecture enables sequential generation of clinically relevant text by predicting the next word based on previous tokens.

3.1.1 | GPT (Generative Pre-Trained Transformer)

The GPT model is an example of a deep-learning model that provides highly qualitative output in terms of language by undergoing pre-training from possibly the largest text data available through self-supervision techniques [25]. In 2017, Google developed a transformer, and then in 2018, OpenAI brought the Generative Pre-trained Transformer into the world [26]. After the initial release, OpenAI has extended a series of variation models, namely GPT-n, which includes GPT-2, GPT-3, and GPT-4 [27]. The most recent advancements in GPT model research are a result of ongoing improvements to the model's design, increased computer power, and the development of novel techniques for task-specific model optimization [28]. Figure 3 illustrates the architecture of the Generative Pre-trained Transformer. Based on the general Feed-Forward Layer in the Transformer architecture, the specific Location-wise Feed-Forward Layer is a fully connected neural network applied to each location in the sequence. It is used in both encoder and decoder stacks, and it's a function that contains two linear transformations with the ReLU activation of various weights and biases for the two distinct layers. This layer is performed after the self-attention in every encoder and decoder block and the absence of labeled data, GPT uses autoregressive language modeling and predicts the coming words by the preceding words [29].

3.1.2 | BERT (Bidirectional Encoder Representations From Transformers)

BERT has enhanced the traditional pre-trained models using bidirectional deep transformers compared to GPT. In this way, it has pre-training and fine-tuning steps using auto-encoding language modeling with masked language modeling (MLM) to be able to detect masked words in contexts in and out of their essays [14]. The BERT algorithm appears flexible as the fine-tuning stage of the NLP tasks can be in its instance. In the pre-training step, it learns the relationships between words in a text using two training objectives: masked language model and next sentence prediction which has been discussed. BERT's architecture (Figure 4) is designed around two key stages: pre-training and fine-tuning [30]. During pre-training, BERT adopts the auto-encoding strategy which has the main objectives of Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) [31]. In the MLM step, certain words in a given sentence are replaced with a specific token and the model learns to predict the replace word from the left and right words in the sentence. This makes it possible for BERT to learn relations between every word sequentially [32]. After completing the pre-training, BERT enters the fine-tuning stage where it learns about particular NLP tasks like sentiment analysis, and question-answering named entity recognition [33]. This phase also enables BERT to get the general focus of the pre-trained parameters and task-specific information much better.

3.1.3 | XLNet

XLNet is a SOTA autoregressive language model [34] that enhances previous transformer-based models, such as BERT, through a unique combination of permutation language modeling and long-term dependency handling [35, 36]. The model demonstrates significant improvements over earlier models in several natural language processing tasks. A key feature of XLNet is its autoregressive modeling with a permutation strategy. Unlike traditional autoregressive models [37, 38], which predict tokens in a fixed left-to-right or right-to-left sequence, XLNet employs a permutation-based training strategy. It permutes the tokens within a sequence, allowing the model to predict the next token using the previous ones. Although the sentence is fed into the model in the correct order, the masking mechanism hides tokens in a permutation-specific way [39]. This ensures that XLNet can leverage bidirectional context without relying on traditional token masking, as seen in BERT. Figure 5 demonstrates the process of XLNet, using the sentence "The sky is blue". The main idea is to model language autoregressively like GPT models but allow for all possible permutations of a sentence. By considering multiple permutations, XLNet captures long-range dependencies and models the bidirectional context of words more effectively. Furthermore, it offers the ability to process complex token dependencies, extending beyond the limitations of traditional models. This is exemplified in cases where XLNet generates predictions based on randomly sampled token orders, thus enhancing its ability to learn deeper contextual relationships within text sequences. Another significant advantage of XLNet is its handling of sequence length. It is one of the few models without a sequence length limitation, providing the flexibility to process very long text sequences.

This model also incorporates the Transformer-XL [40] backbone, which employs a recurrence mechanism to build long-term dependencies between tokens. This mechanism allows the model to capture information across long text sequences effectively, making it particularly suited for tasks that require extended context. Moreover, XLNet eliminates the need for masking tokens during training, unlike BERT's masked language modeling

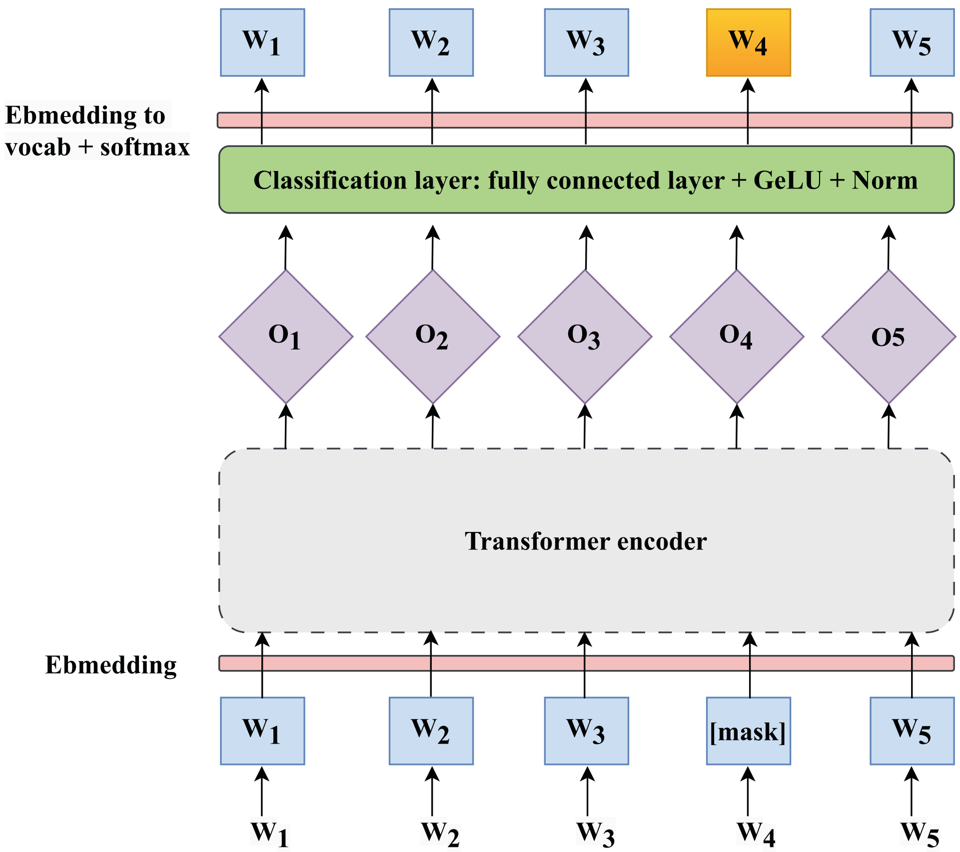

FIGURE 4 | Visualization of BERT's architecture, which uses masked language modeling to learn bidirectional representations of text. This is particularly effective in extracting contextual meaning from medical documents, aiding tasks like question answering and clinical summarization.

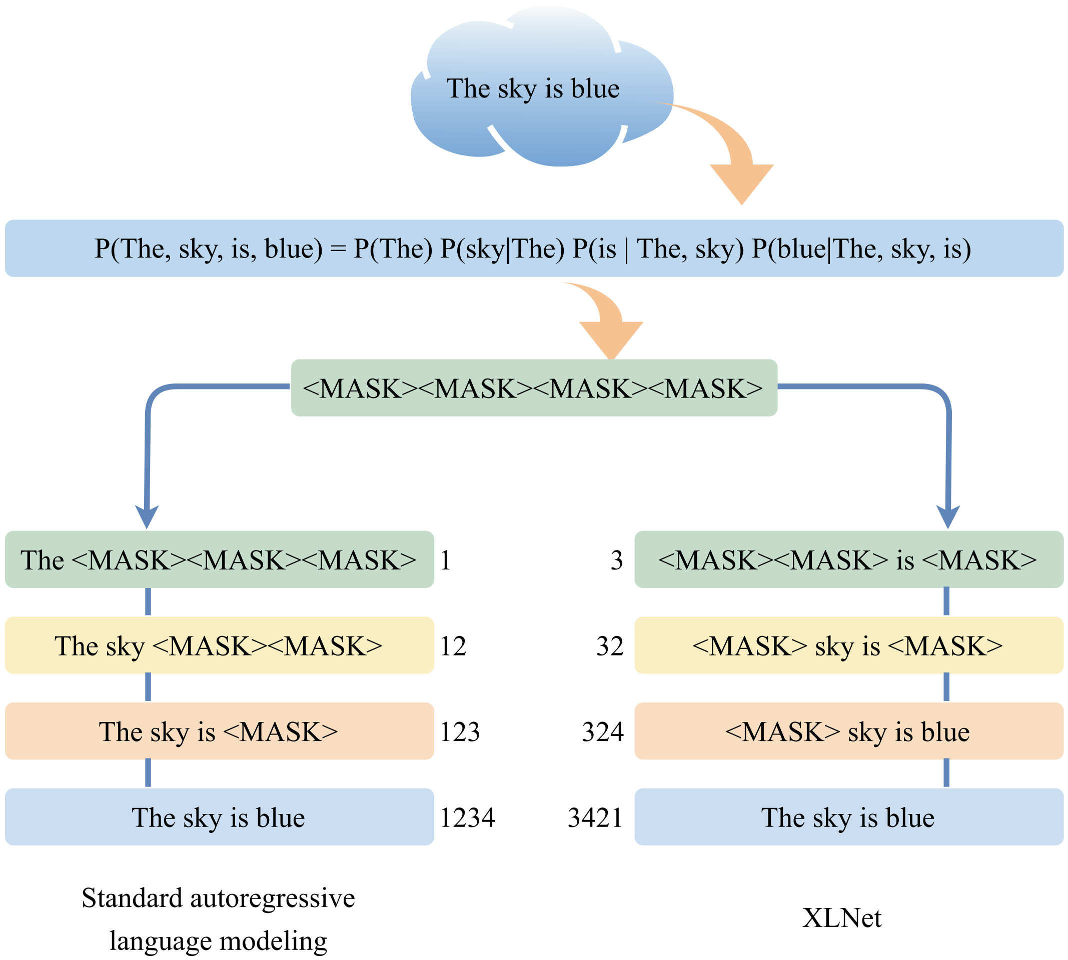

FIGURE 5 | Illustration of XLNet's permutation-based training and autoregressive modeling techniques. Using a sample sentence and permutation logic, this figure highlights how XLNet improves contextual understanding—crucial for robust performance in complex clinical language settings.

(MLM) [41], which results in a more seamless alignment between training and inference. By integrating autoregressive modeling with a permutation strategy and long-range dependency handling, XLNet achieves superior performance across a variety of natural language processing tasks, including text classification, question answering, and language modeling [35]. Its ability to model complex token dependencies without a sequence length limit gives it a substantial advantage over the previous models.

3.1.4 | RoBERTa

RoBERTa, an acronym for "Robustly Optimized BERT Pretraining Approach," is an advanced NLP model introduced by Facebook AI in 2019 which is built upon the foundation laid by Google's BERT [42]. While BERT revolutionized NLP by enabling deep bidirectional contextual learning, it refines this model with key optimizations that address limitations in BERT's pretraining and fine-tuning procedures. As a result, this model delivers superior performance across a wide range of NLP benchmarks, including text classification, machine translation, and question-answering tasks. RoBERTa incorporates several crucial enhancements over the original BERT model, including training on a significantly larger data set and eliminating the Next Sentence Prediction (NSP) task. The larger training corpus allows it to capture more linguistic diversity and better generalize across tasks. The NSP objective was found to add unnecessary noise, so this model focuses solely on masked language modeling (MLM), which improves training efficiency [43]. Furthermore, RoBERTa uses dynamic masking, wherein tokens are randomly masked at each training epoch, preventing overfitting and promoting a deeper understanding of diverse contexts [44]. It also employs a byte-level Byte Pair Encoding (BPE) tokenizer, allowing more effective handling of rare and out-of-vocabulary words. These improvements collectively enable it to outperform BERT in various natural language tasks [45]. Additionally, RoBERTa's performance gains are comparable to other transformer-based innovations like ALBERT, a lightweight version of BERT designed for efficiency [46], and the T5 (Text-to-Text Transfer Transformer) model, which emphasizes unified text processing across NLP tasks [47]. The enhancements introduced in RoBERTa have led to substantial improvements in model performance across numerous NLP benchmarks. RoBERTa achieved SOTA results on several key tasks, including the General Language Understanding Evaluation (GLUE) benchmark [48], SQuAD (Stanford Question Answering Dataset) [49], and RACE (Reading Comprehension data set for English) [50]. By eliminating the next sentence prediction task and increasing training data, RoBERTa's robust pretraining approach allows it to outperform BERT in most cases, demonstrating its superior capacity for language understanding and transfer learning.

3.1.5 | ALBERT

ALBERT (A Lite BERT) [51] is a lightweight version of BERT, developed by Google in 2019 for natural language processing tasks which can also be considered part of the broader family of Large Language Models since it is built on the transformer architecture and trained on large-scale language data. It employs parameter-reduction techniques to improve the efficiency and performance of the model while retaining much of BERT's language understanding capabilities. ALBERT reduces parameters by employing cross-layer parameter sharing [52] and factorized embedding parameterization, making it much lighter (18M parameters compared to BERT's 108M) and faster in training and inference. It decouples the embedding size from the hidden layer size, reducing the size of the vocabulary embeddings, one of BERT's largest parameter-heavy components. This further contributes to lower the overall model size. ALBERT also introduces the Sentence Order [53] prediction task, where the model is trained to distinguish if two sentences appear in the correct order or are swapped. This task improves coherence and understanding of text, which is particularly useful for tasks such as clinical document analysis and medical literature summarization. Figure 6 illustrates the basic structure of ALBERT using a transformer.

In medical applications, the computational resources are limited, such as mobile health monitoring, ALBERT's smaller size, and faster performance making it an ideal choice for tasks like clinical document classification, medical summarization [54], and predictive modeling. Its efficiency enables seamless integration into healthcare systems for real-time disease detection and symptom checking, delivering quick and accurate responses in environments with restricted processing power, such as telehealth platforms and remote diagnostics. ALBERT's capacity to efficiently process and summarize medical literature and clinical documentation with reduced computational resources makes it particularly well-suited for the automated summarization of Electronic Health Records (EHRs) [55]. This enables healthcare providers to navigate large volumes of data more effectively. Additionally, ALBERT can be integrated into clinical decision support systems to analyze medical literature and generate recommendations or predictive insights, enhancing diagnostic precision in areas, such as radiology [56] and pathology [57].

ALBERT's reduced architectural complexity and enhanced computational efficiency make it a highly valuable tool for deploying Large Language Models in medical applications. Its adaptability to a range of healthcare tasks, combined with its accelerated training and inference capabilities, positions it as an effective model for digital health technologies and clinical decision-making systems. By enabling resource-efficient processing without sacrificing performance, ALBERT offers significant potential for advancing diagnostic accuracy and operational efficiency in various medical and digital health contexts. Although ALBERT has a smaller memory footprint due to its repeating layers, its computational cost is similar to that of traditional BERT-like architectures.

3.1.6 | Megatron

To increase the trainability of NLP tasks, NVIDIA developed the Megatron Large Language Model [58]. Similar to BERT and GPT, it draws upon the Transformer architecture, (Figure 7) but it is designed to scale up training across several GPUs effectively. Megatron's primary characteristic is its large model sizes, which have billions of parameters and allow it to comprehend intricate linguistic patterns [59]. The parallel model can effectively handle massive computational tasks since it is specifically made to be taught distributedly over thousands of GPUs. Additionally, Megatron uses techniques like mixed-precision training to advance training while maintaining an intact accuracy rate [39]. The main attractive feature of this model is the ability to perform different

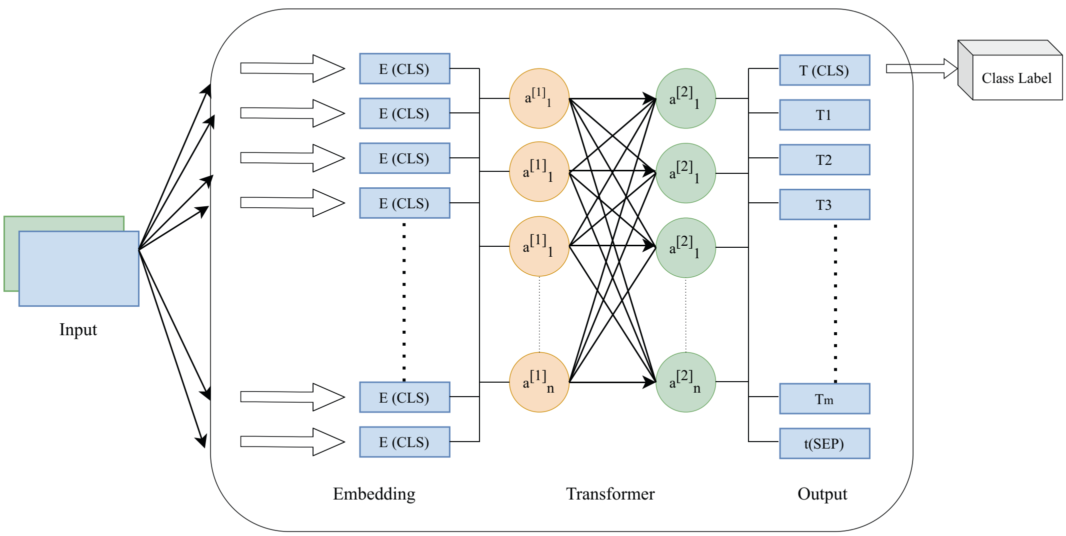

FIGURE 6 | Depiction of ALBERT's architecture, optimized for efficiency by employing parameter-sharing and factorized embeddings. This lightweight model structure supports resource-constrained healthcare applications while maintaining competitive performance in natural language tasks.

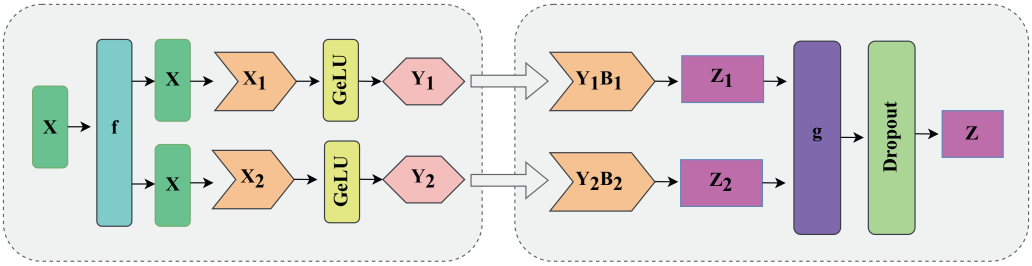

FIGURE 7 | Visualization of Megatron-LM's scalable architecture that employs model parallelism for training large-scale LLMs across multiple GPUs. This capability is essential for developing highly expressive models suitable for medical language tasks involving vast data sets.

NLP tasks like text generation, machine translation, and question answering that result from the optimized language understanding [60]. Megatron's ability to read through millions of data inputs at once makes it the perfect tool for sectors like filtering content, the legal department, and the healthcare industry, which handles a lot of text communications. In real life, this LLM can be used to address a variety of NLP tasks, including the analysis of changes in fine-grained textual content, closely examining changes in language pauses, and more detailed contextual analysis.

3.2 | Generative Models

Generative models are a separate class of machine-learning algorithms that are intended to generate fresh data samples that mirror the properties of preexisting data sets. These are useful for creating synthetic data in industries like healthcare and finance where real data may be limited since they capture and repeat data patterns, in contrast to traditional models that concentrate on classification or prediction. By simulating real-world situations, this synthetic data aids in research and model training and improves the precision of data-driven insights [61]. Additionally, generative models are widely utilized in the creative industries to produce literature, artwork, and multimedia material, making the process of creating content for virtual worlds and interactive experiences more efficient [62]. The main goal of generative modeling is learning a probability distribution over input data, which makes it possible to create fresh representative samples. Several strategies are employed to do this [63]. For example, autoregressive models [64] are excellent at generating language because they generate text sequentially, building on previous words to get outputs that are logical and pertinent to the context. Long-range dependencies in language are captured by this sequential capacity, which is frequently implemented using transformer architectures [27]. Variational Autoencoders (VAEs) [65] is an additional method that encodes data into a latent space, enabling the controlled creation of new samples. VAEs are perfect for jobs requiring output variance, including producing a variety of images or targeted synthetic data, because of their compact representation.

The main loss function for a Variational Autoencoder (VAE) combines [66] a reconstruction term with a regularization term to enforce structured latent space. This loss function helps the VAE learn both a meaningful latent space and accurate reconstructions of the input data. Given an input x and its latent representation z, the VAE loss $\mathcal{L}_{\text{VAE}}$ is defined in Equation (1).

$$\mathcal{L}_{\text{VAE}} = D_{\text{KL}}\left(q_{\phi}(z|x) \parallel p(z)\right) - \mathbb{E}_{q_{\phi}(z|x)}\left[\log p_{\theta}(x|z)\right] \tag{1}$$

where the Kullback–Leibler divergence $D_{\text{KL}}\left(q_{\phi}(z|x) \parallel p(z)\right)$ term encourages the latent variable z to follow a prior distribution, often Gaussian, while the reconstruction loss $\mathbb{E}_{q_{\phi}(z|x)}\left[\log p_{\theta}(x|z)\right]$ measures the difference between the original input x and its reconstruction $\hat{x}$.

Generative Adversarial Networks (GANs) [67] represent a breakthrough in generative modeling, employing an adversarial configuration with two neural networks: a discriminator and a generator. The discriminator determines if the generated data is authentic or not. This dynamic forces the generator to advance, producing synthetic outputs of superior quality. Applications in visual media, particularly the creation of realistic images, films, and intricate digital content, have been significantly impacted by GANs. Here are some of the primary loss functions [68] used in GANs:

(i) The adversarial loss: Adversarial loss is central to the GAN framework. It governs the interaction between the generator F and the discriminator T. The discriminator T learns to distinguish real images e from generated images $F(v)$, while the generator F strives to produce images that can deceive the discriminator. The adversarial loss is mathematically expressed in Equation (2).

$$\mathcal{L}_{\text{GAN}}(F, T) = \mathbb{E}_{e}[\log T(e)] + \mathbb{E}_{v}[\log(1 - T(F(v)))] \tag{2}$$

(ii) The content loss: Content loss or reconstruction loss measures feature-level similarity between a source image q and a generated image r by using a pre-trained network's feature representations at a specific layer. This loss function encourages the generated image to maintain the structural details of the source image. Given feature representations Q^{j}_{mn} and R^{j}_{mn} for images q and r at layer j, the content loss is defined in Equation (3).

$$\mathcal{L}_{\text{content}}(q, r, j) = \frac{1}{2}\sum_{m,n}(Q^{j}_{mn} - R^{j}_{mn})^2 \tag{3}$$

(iii) The L_1 loss: This loss focuses on minimizing the absolute pixel-wise difference between a target image s and a generated image u. It captures low-frequency details, enforcing structural similarity between the generated and target images. This loss function is commonly used to promote accurate reconstruction of the target image's overall appearance and details. The L_1 loss is expressed in Equation (4).

$$\mathcal{L}_{L_1}(u) = \mathbb{E}_{a,s}\left[\|s - u\|_1\right] \tag{4}$$

Applications of generative models are found in many different fields. In health care, they create artificial patient data that preserves important trends while safeguarding privacy, facilitating model building and study without disclosing private information. Generative models support advancements in digital art and virtual reality as well as drive innovation in digital content in the creative sectors. Larger data sets and more complicated structures can be handled by these models as they become more effective and scalable, which further establishes their place in artificial intelligence research and applications.

3.3 | Autoregressive Models

Autoregressive (AR) models are statistical techniques widely applied in time series forecasting, where the current value of a variable is predicted using a weighted sum of its prior values, along with a stochastic component [69]. These models represent a specialized form of regression analysis that emphasizes temporal dependencies by regressing a variable onto its lagged (previous) values [70]. Initial parameter estimation for an AR model is critical and often carried out through methods such as the Yule-Walker equations [71], Maximum Likelihood Estimation (MLE), or Least Squares Estimation [72]. Upon estimating the parameters, the model becomes capable of forecasting future values within a time series. Thereby, the model can capture patterns that traditional regression models for cross-sectional data cannot. Mathematically, an autoregressive model of order p can be expressed in Equation (5).

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{5}$$

where y_t denotes the value at time t, $\phi_1, \phi_2, \ldots, \phi_p$ are autoregressive coefficients, p is the model order, and ε_t represents the error term (white noise) at time t.

In the realm of large language models, autoregressive models are instrumental in generating sequences of text by predicting each token based on prior tokens in the sequence. This approach enables LLMs to create coherent, contextually relevant text by statistically conditioning each output on its preceding outputs. Employing transformer architectures, autoregressive language models execute complex computations to produce language that mirrors human-like conversation, with the capacity to anticipate subsequent text or behavior based on historical data [73]. These models, often incorporating billions of parameters, are proficient at handling vast text data and make them invaluable in applications requiring substantial contextual understanding and text generation abilities.

3.4 | Natural Language Processing

Natural Language Processing (NLP) is a vital area in artificial intelligence (AI) and computer science, aiming to bridge the gap between human communication and machine understanding [74]. It combines theories and methods from computer science, linguistics, and mathematics to interpret and execute human language commands as depicted in Figure 8. NLP encompasses two major research directions: Natural Language Understanding (NLU) and Natural Language Generation (NLG) [75]. NLU focuses on enabling machines to comprehend and analyze natural language by identifying entities, emotions, concepts, and keywords [76]. It is widely used in applications like customer service, where machines interpret user-reported problems either through text or speech. NLU places a major priority on language study, which focuses on language's meaning and context [77]. In contrast, NLG is used to create a human-like language from internal data and is composed of activities such as goals, planning, and text realization to generate sentences or paragraphs [75]. Regarding the process of reducing computational complexity and improving the efficiency of performing text-based language processing, the specific algorithms [78] include Long Short-Term Memory (LSTM) [79], Sequence-to-Sequence [80], Named Entity Recognition (NER), User Preference Graph, Word Embedding [81], and so on. These models help to solve such problems as sentence extraction, automatic text summarization, fuzzy inference

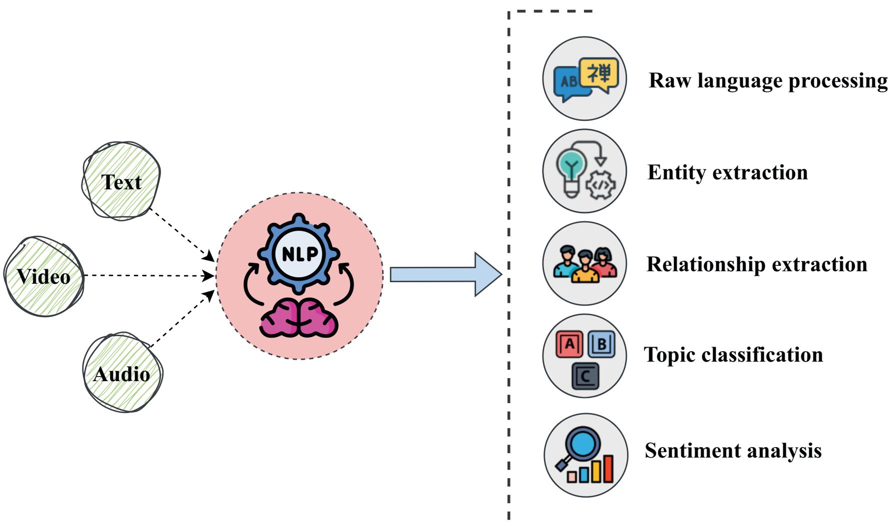

FIGURE 8 | Workflow diagram of key NLP stages, from data ingestion and preprocessing to model training and output generation. This pipeline forms the backbone of LLM-driven medical applications such as diagnostic reasoning, summarization, and automated documentation.

rule-based processing, and, as a result, help to create machines that better understand human language. NLP is essential to the healthcare industry because it can convert vast amounts of unstructured data in electronic health records into forms that are helpful for diagnosis, treatment planning, and decision-making [82]. NLP also comes in handy in predictive analytics in the sense of diagnosing disease onset through symptoms, history, and lab test results.

4 | Taxonomy of Large Language Model (LLM) Applications in Medical and Digital Health

The taxonomy categorizes the wide range of applications for medical LLMs, showcasing their versatility in both clinical and research settings. These LLMs Model (Figure 9) improve diagnostic and predictive capabilities, streamline patient interactions, enhance drug discovery, and contribute to public health efforts, marking a new era of AI-driven medical advancements. Large Language Models, in conjunction with Natural Language Processing, play an increasingly critical role in transforming the healthcare landscape, with a broad spectrum of applications. In the domain of clinical decision support, LLMs provide substantial benefits by helping healthcare providers make more informed decisions.

LLMs analyze vast quantities of medical literature, patient histories, and real-time clinical data, offering diagnostic assistance that aids physicians in identifying diseases and interpreting symptoms. Furthermore, LLMs are used to build predictive models, which can forecast disease progression, predict patient outcomes, and suggest optimal treatment paths based on historical data. Another major area of application is medical text processing and summarization. LLMs are employed to streamline administrative burdens by automating clinical documentation, allowing physicians to focus more on patient care. By parsing Electronic Health Records (EHRs), LLMs can extract critical patient information, offering real-time insights during clinical encounters. Additionally, LLMs are invaluable in medical literature synthesis, helping clinicians stay updated by summarizing complex research papers, clinical guidelines, and systematic reviews into concise, actionable insights. LLMs also play a pivotal role in enhancing patient interaction and support. Through virtual health assistants and chatbots, LLMs provide patients with 24/7 access to healthcare guidance, answering common medical questions and offering advice based on symptom reporting. These models power symptom checkers that use patient input to suggest possible conditions, providing preliminary assessments that can guide individuals on whether to seek medical help. By utilizing natural language understanding, healthcare chatbots can engage in smooth, human-like conversations, significantly enhancing the patient experience. In the area of drug discovery and pharmacology, LLMs accelerate the identification of potential drug interactions through text mining, sifting through vast data sets of biomedical information. They also play a crucial role in drug discovery, where LLMs analyze chemical compositions and biological pathways, speeding up the development of new therapies. Moreover, LLMs contribute to disease detection, particularly in

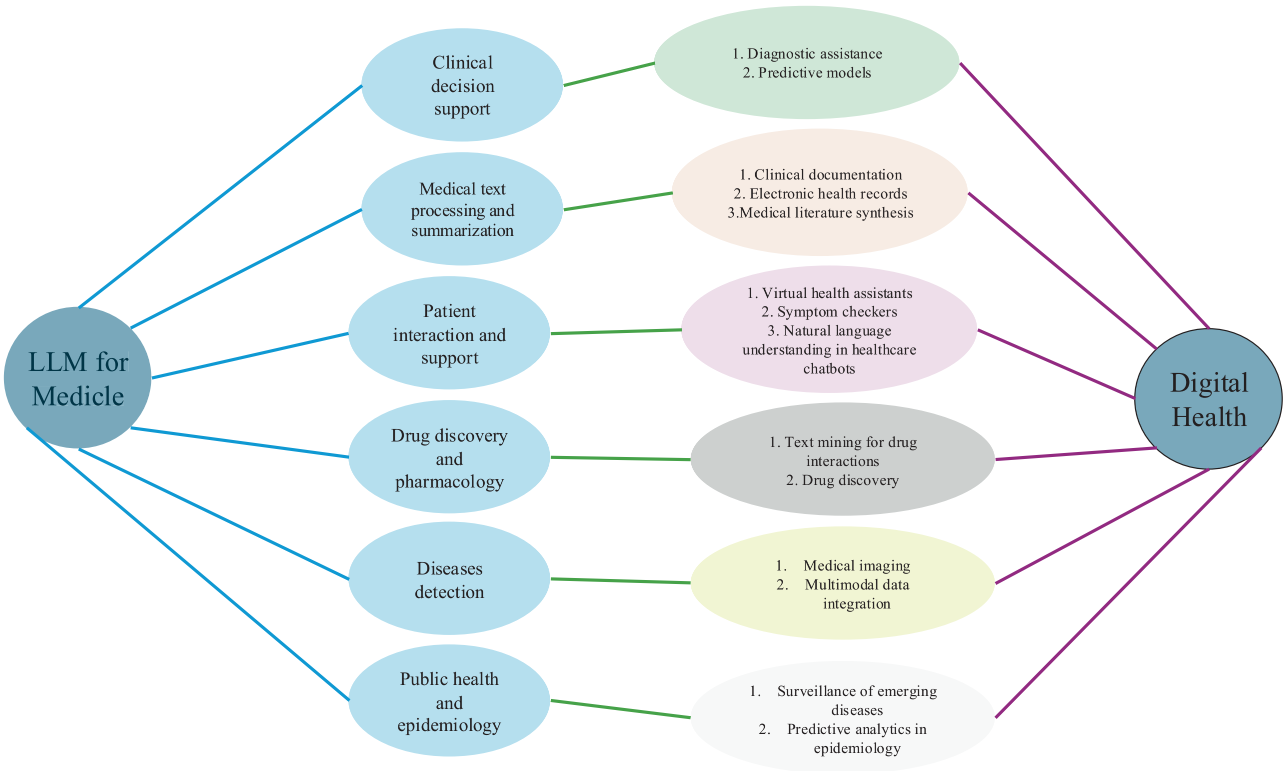

FIGURE 9 | This figure presents a structured taxonomy categorizing the diverse applications of LLMs across key domains in medical and digital health. It classifies use cases such as clinical decision support, medical documentation, drug discovery, patient interaction, and public health monitoring. The taxonomy provides a comprehensive overview that helps researchers and practitioners understand the functional landscape of LLM integration in clinical practice.

medical imaging analysis, where they assist in interpreting complex imaging data. By integrating multimodal data, LLMs provide comprehensive diagnostics by combining data from various sources like imaging, lab results, and genetic information.

Beyond individual patient care, LLMs have broad implications for public health and epidemiology. They aid in disease surveillance, tracking emerging infectious diseases, and offering predictive analytics to anticipate outbreaks or monitor public health trends. Despite the vast potential of LLMs, their implementation faces challenges such as ensuring data privacy, improving interpretability, and overcoming barriers to deployment in clinical settings. Nonetheless, the promise of LLMs to revolutionize medical practices, from patient care to drug development, marks a significant leap forward in healthcare innovation.

5 | LLM Applications in Medical and Digital Health

Large Language Models are increasingly transforming the medical and biomedical fields by advancing diagnostics, enhancing patient engagement, and streamlining clinical workflows [83, 84]. LLMs provide clinicians with rapid and accurate extraction of complex information from electronic health records (EHRs), including comprehensive patient histories, treatment outcomes, and relevant medical literature, thereby optimizing clinical decision-making [85]. LLMs play a crucial role in personalized health care by analyzing patient records and real-time health data to provide customized recommendations for nursing care. That also predicts disease progression and fosters coherence in healthcare delivery [86]. By synthesizing vast medical literature, these models support evidence-based practices across healthcare sectors beyond clinical applications [87]. LLMs facilitate complex diagnostics and treatment planning by integrating and interpreting data sources, such as laboratory results, imaging findings, and genetic information. This integration enables clinicians to make informed, data-driven decisions and enhances diagnostic accuracy. To identify and track emerging health trends, LLMs contribute to disease surveillance and predictive analytics by examining diverse data, including public health records, social media patterns, and epidemiological reports. LLMs enhance innovation by helping to discover new drug targets, comprehend proteinŰligand interactions, and forecast adverse effects or interactions between drugs. Large language models simplify patient education by generating explanations of medical conditions, treatment strategies, and lifestyle changes, which enhance health literacy and patient management. These capabilities are particularly impactful within healthcare chatbot systems, where LLMs deliver immediate, evidence-based information on symptoms, medications, and self-care practices. The adaptability and wide range of applications of LLMs make them a valuable asset in bridging healthcare delivery gaps across diverse clinical settings.

5.1 | Clinical Decision Support

Clinical Decision Support Systems (CDSS) are advanced software tools designed to support clinicians in making evidence-based medical decisions, aiming to elevate healthcare quality and patient outcomes. CDSS systems integrate health IT infrastructures to deliver timely, data-driven insights to support the entire care team, including the patient [88]. Key channels, such as Electronic Health Records and Health Management Systems, facilitate information delivery, while structured intervention formats such as order sets, flow sheets, patient lists, and dashboardsŮaid in precise decision-making processes [89]. CDSS fosters a more cost-effective and efficient clinical environment by reducing misdiagnoses and medication errors. Recent advancements in LLMs have significantly enhanced the capabilities of clinical decision support systems by reinforcing their essential role in modern healthcare settings. The application of state-of-the-art LLMs, such as OpenAI's GPT-3, has led to improvements in accuracy, consistency, and clinical efficacy within CDSS frameworks, solidifying their role in modern healthcare settings [90].

Building upon these advancements, Med-PaLM developed using Google's PaLM architecture demonstrates refined clinical decision support through advanced tuning of language models. This provides evidence-based recommendations that enhance diagnostic accuracy and elevate patient care to meet rigorous healthcare standards [91]. Figure 10 depicts the architecture of CDSS employing large language models or natural language processing. It gathers and processes diverse patient data, including laboratory test results, monitoring data, medical history, symptoms, and demographic information. This data is transformed into integrated features for input into the Clinical Decision Support Module. Here, LLMS or NLP algorithms analyze the combined data and produce insights to guide clinical decisions which provides the clinician with recommendations that aid in patient management. This architecture emphasizes how LLMs can streamline complex data handling, automating the analysis of large data sets to improve diagnostic accuracy and decision support in healthcare settings. By effectively synthesizing information from multiple sources, LLM-powered CDSS can help clinicians make informed decisions, improve outcomes, and optimize the healthcare process.

5.1.1 | Diagnostic Assistance

Diagnostic assistance within CDSS empowers clinicians by leveraging predictive analytics and extensive medical knowledge bases to enhance diagnostic accuracy and efficacy. Diagnostic Decision Support Systems (DDSS) are specialized CDSS tools tailored to assist in clinical diagnostics [92, 93]. These systems incorporate computerized consultation and filtering mechanisms, where data inputs are systematically analyzed to produce probable diagnoses [94]. Advanced methodologies, including artificial neural networks and deep learning, play a pivotal role in supporting disease diagnosis and treatment, thus enabling healthcare professionals to make more informed decisions [95, 96]. By systematically assessing multiple diagnostic possibilities, DDSS tools serve to reduce diagnostic errors and enhance clinical outcomes. To support more rigorous evaluation, a benchmarking data set focused on anesthesiology was recently introduced, designed to evaluate the diagnostic performance of medical LLMs across specialty-specific tasks using structured, domain-adapted metrics [97].

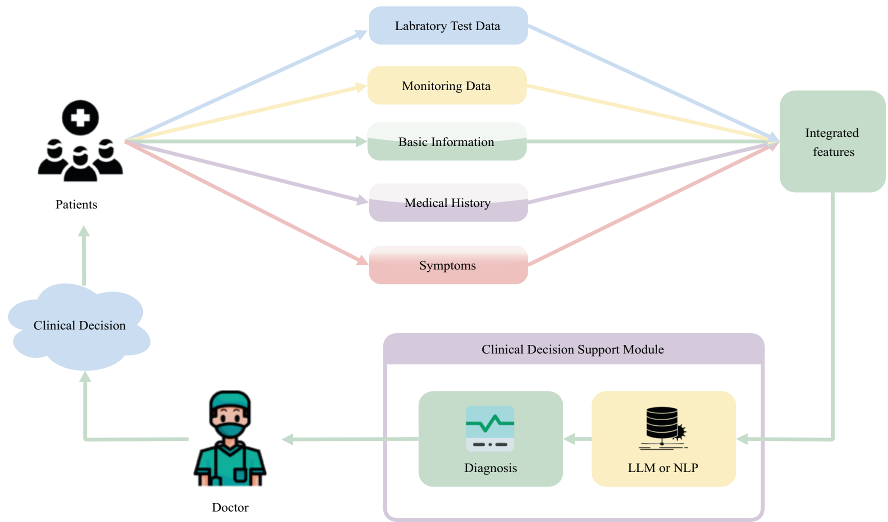

FIGURE 10 | Schematic diagram of a Clinical Decision Support System (CDSS) integrated with LLM or NLP modules. The system is designed to analyze clinical data inputs and assist healthcare professionals by delivering evidence-based recommendations and alerts.

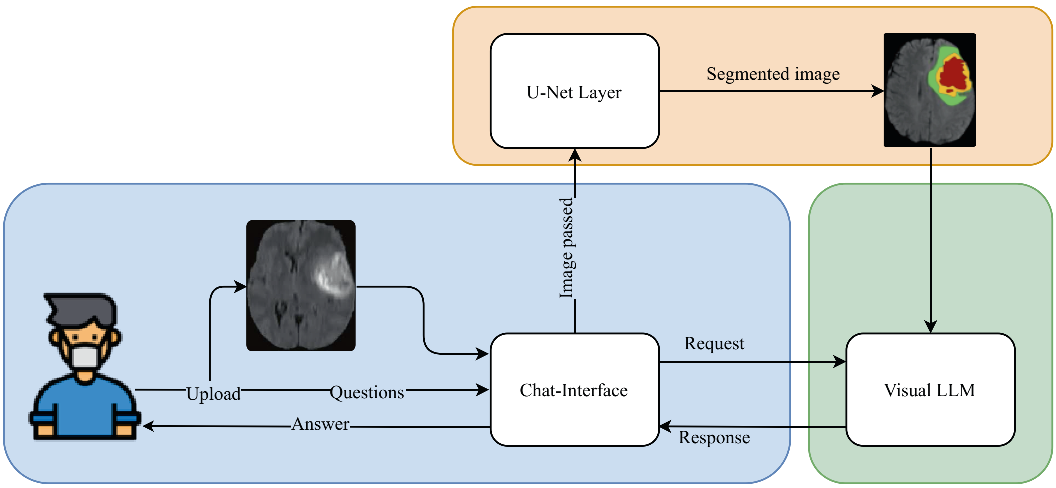

FIGURE 11 | Depiction of a hybrid diagnostic framework combining Visual LLM and U-Net segmentation. This model aids clinicians by extracting anatomical features from medical images and answering.

Figure 11 illustrates a diagnostic assistance framework that integrates a Visual LLM [98] and U-Net Layer [99] to support clinicians in medical image analysis and decision-making. In this workflow, a clinician uploads a medical image, which is processed through the U-Net Layer to produce a segmented image, highlighting key regions, such as potential abnormalities. The segmented image is subsequently analyzed by the Visual LLM, which uses the segmentation data to generate responses to diagnostic queries. Through a Chat Interface, the clinician can interact with the system by asking specific diagnostic questions and receiving precise, contextually relevant answers [100, 101]. This framework exemplifies the potential of LLMs to deliver targeted diagnostic support by integrating image processing with language-based interaction, offering an intuitive, interactive tool for clinicians to derive insights from complex imaging data.

5.1.2 | Predictive Models

Predictive models in healthcare leverage data-driven methodologies to forecast patient outcomes, support diagnostic processes, and guide clinical decision-making by identifying intricate patterns within vast datasets. These models serve as essential components of clinical decision support systems (CDSS), facilitating early disease detection, interpretation of radiology reports, and analysis of pathology slides. Predictive models can analyze imaging data and pathology slides to detect signs of disease at an early stage [102]. It can also assist radiologists by identifying anomalies in X-rays, MRIs, and CT scans to help the early detection of conditions like tumors, fractures, and pulmonary diseases [103]. Clinicians may anticipate health risks and customize care strategies by using predictive models to forecast patient readmission risks, mortality rates, and probable illness consequences. It can detect patients who are at an increased risk of being readmitted by examining factors such as their prior hospital stays, current health issues, medications, and social determinants of health [104]. This enables healthcare professionals to take a proactive approach to manage these patients, decrease rates of readmission, and enhance health outcomes.

Effective management of risk can be mitigated by utilizing AI-driven predictive analytics and Natural Language Processing [105]. By integrating diverse data sources, including electronic health records, laboratory findings, and demographic information, predictive models offer comprehensive insights that help clinicians prioritize care and make informed treatment decisions, specifically in critical care settings [106]. Recent advancements in LLMs have significantly augmented predictive modeling capabilities in health care by automating feature selection, improving interpretability, and supporting complex clinical decisions. LLMs can efficiently identify high-impact features for healthcare outcomes and analyze EHR data to uncover crucial predictive indicators without relying heavily on labeled data, thereby streamlining the data preparation process [107]. Furthermore, by transforming structured patient data into narrative formats, LLMs enhance the precision of disease prediction through few-shot learning techniques, which allow these models to generalize from limited samples of information [108].

5.1.3 | Medication Recommendation

In recent years, the application of Natural Language Processing and Large Language Models in medication recommendation has advanced significantly, targeting improvements in personalized treatment plans, patient safety, and the reduction of medication errors. By analyzing diverse clinical data, including patient histories, Electronic Health Records (EHRs), and medical literature, NLP algorithms, and LLMs can recommend appropriate medications, enhancing decision-making in clinical settings. ShennongMGS, for instance, is an LLM-based Chinese medication guidance system designed to address polypharmacy and lagging data challenges. It improves medication guidance and predicts adverse drug reactions by simulating pharmacists' decision-making and incorporating self-updating knowledge, elevating drug safety and healthcare quality [109].

Building on this momentum, medical LLMs have gained traction as powerful tools in health care, where their applications span medication guidance and adverse reaction prediction. Despite their promise, LLMs encounter challenges in addressing the complexity of polypharmacy and real-time updates. A study reviewing the development of medical LLMs highlights these concerns and underscores the need for advancing research in this domain. The study examines the core requirements for building effective medical LLMs and their potential to revolutionize health care, particularly in clinical settings, where these models could offer enhanced medication recommendations and patient safety measures [110]. Simultaneously, efforts to integrate traditional medicine into these frameworks have also shown promising results. For example, advancements in Ayurveda literature digitization have led to a centralized platform that processes Ayurvedic data, aiding in drug recommendations. An Ayurvedic drug recommendation model achieved impressive accuracy and response time, demonstrating that even non-Western medicine systems can benefit from the scalability and reliability of AI-driven platforms [111]. This model could serve as a template for integrating alternative medicine systems into the broader landscape of LLM-based medical guidance.

The Covid-19 pandemic further amplified the urgency of machine- learning applications in medication management. As global healthcare systems faced unprecedented medication shortages, machine-learning models like the a priori algorithm were proposed to help manage pharmaceutical stock more effectively. This approach ensures that high-demand medications remain accessible, thereby preventing shortages and enhancing the management of limited healthcare resources during crises [112]. To reduce the cognitive load on healthcare professionals, AI-driven systems such as the Artificial Intelligence-based Neural Drug Suggestion Model (AINDSM) have been developed. This system significantly lowers the workload by providing automated, efficient drug recommendations. When evaluated against traditional models like Support Vector Machines (SVMs), AINDSM demonstrated improved performance, offering the potential to streamline drug recommendation processes and improve the overall efficiency of clinical decision-making [113].

The use of LLMs for medical content generation has also expanded into specific clinical specialties such as ophthalmology, orthopedics, and dermatology. A study evaluating four prominent LLMsŮClaude-instant-v1.0, GPT-3.5-Turbo, Command-large-nightly, and BloomzŮhighlighted how these models could generate relevant medical content, offering clinicians additional tools for patient care. While promising, these LLMs still face limitations in accuracy and application, requiring further refinement before being fully integrated into clinical workflows [114]. ShennongGPT represents another advanced LLM designed for precise medication guidance and adverse drug reaction predictions. This model utilizes a two-stage training approach, combining foundational knowledge from distilled drug databases with real-world patient data to simulate human reasoning. This dual training method enhances the model's applicability in clinical settings, providing a more comprehensive and relevant medication recommendation system that adapts to real-time patient needs [115].

Ensuring clear and accurate communication of medication directions remains a priority, particularly in pharmacy practice, where errors in dosage or frequency instructions can have severe consequences. The MEDIC system, which combines domain knowledge with LLMs, aims to minimize such errors by emulating pharmacist reasoning. Through this system, key prescription elements like dosage and frequency are conveyed, helping reduce the risk of adverse drug events and improving patient outcomes [116]. The broader application of medical LLMs, while promising, still faces challenges. For example, LLMs often struggle with drug name recognition due to the prevalence of synonyms and can occasionally generate inaccurate or hallucinated information. An initial assessment of LLMs as providers of drug information to patients demonstrated these limitations, highlighting the importance of refining models to improve reliability and accuracy in real-world medical applications [117]. LLMs and AI-driven medication recommendation systems are revolutionizing healthcare by enhancing treatment personalization, reducing medication errors, and improving patient safety. However, challenges such as accuracy, polypharmacy complexity, and real-time adaptability remain, necessitating further refinement. Despite these hurdles, the potential for LLMs to streamline healthcare processes and enhance patient outcomes is evident, with ongoing advancements paving the way for their wider integration into clinical practice.

5.2 | Medical Text Processing and Summarization

In the realm of health care, there is a vast amount of unstructured text data, which presents both challenges and opportunities for improving patient care, clinical decisions, and research. Medical text processing and summarization (Figure 12) have become essential for turning this data into useful information. With advancements in Natural Language Processing and Large Language Models, it is now easier to process and summarize complex medical information. This section looks at how LLMs and NLP are used in key areas like clinical documentation, Electronic Health Records (EHRs), and summarizing medical research, making healthcare information more accessible and efficient.

5.2.1 | Clinical Documentation

Comprehensive clinical documentation is widely recognized as a cornerstone of effective healthcare delivery. Nevertheless, the extensive workload required for accurate documentation frequently results in significant stress and burnout among healthcare professionals. This added pressure not only increases the likelihood of medical errors but also poses serious risks to patient safety, further underscoring the urgent need for supportive tools such as NLP [118]. NLP provides a valuable solution by automating the extraction and organization of information from free-text clinical notes, which effectively alleviates some of the documentation burden on healthcare providers. This benefit is evident in various fields; for instance, in the management of chronic diseases, NLP techniques have demonstrated great utility in analyzing the intricacies of complex clinical narratives. Research in this area highlights how LLM methodologies can be tailored to capture the unique and often complicated linguistic structures found in chronic disease documentation, showcasing the adaptability of LLM in chronic care environments [119]. Similarly, in

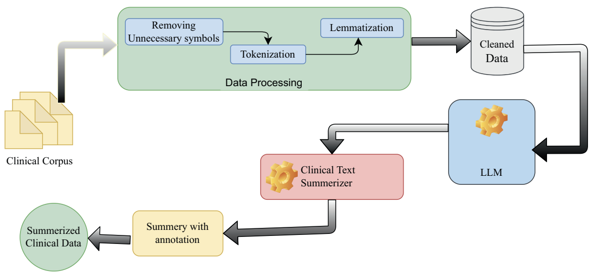

FIGURE 12 | Illustration showing how LLMs process unstructured clinical texts to extract relevant information and generate summaries. This capability is essential for streamlining documentation and supporting physician decision-making.

dentistry, researchers have explored the applications of NLP for information retrieval from clinical notes. This field encounters specialized terminology and nuanced language that present its own set of challenges, further underscoring the potential for NLP methods to adapt across diverse medical specialties [120].

Oncology serves as another compelling example of LLM's impact on clinical documentation, especially in automating the extraction of cancer-related concepts from clinical notes. Studies in this domain highlight the effectiveness of rule-based algorithms, which achieve a high level of accuracy and sensitivity in identifying cancer-specific terms, thereby suggesting a promising future for similar applications in other disease areas [121]. Extending from oncology into the field of nursing, NLP also shows potential in the extraction of symptom-related information, although the complex and varied language used to document symptoms in clinical notes reveals a need for methods specifically designed to handle this aspect. Symptom extraction can facilitate large-scale data reuse and enhance research focused on symptom patterns across patient populations [122]. Another critical application of LLM is constructing clinical registries, which are instrumental in supporting patient care quality improvement. LLM methods have shown promise in extracting structured and valuable data from unstructured clinical notes; however, the distinct nature of clinical texts often diverges from the standard data on which LLM models are trained and poses unique challenges in healthcare applications [123].

As LLM technology continues to evolve and refine its methodologies, its role in healthcare communication is becoming increasingly significant. By facilitating clearer information exchange and better understanding between patients and healthcare providers, LLM plays a direct role in enhancing patient outcomes, encouraging shared decision-making, and supporting collaborative care processes [124]. Beyond facilitating communication, LLM offers critical support in the codification of medical diagnoses and procedures. This task is essential for healthcare systems that rely on efficient and accurate coding to manage resources effectively and ensure proper reimbursement. LLM's ability to automate traditionally manual aspects of this process reduces the likelihood of errors and contributes to improved accuracy in medical code assignments [125]. The versatility of NLP is further illustrated through its applications in specific medical contexts. For example, in preoperative settings, a clinical NLP pipeline has been used to extract elements from medical history within clinical notes, demonstrating its utility in focused clinical scenarios [126].

Collectively, these studies emphasize the expanding role of LLMs in alleviating documentation burdens, supporting research across diverse medical fields, and improving patient care. They also underscore the need for continuous advancements to tackle the unique linguistic complexities embedded in clinical notes, especially as these models are applied to the specialized and varied domains within health care.

5.2.2 | Electronic Health Records (EHRs)

Electronic Health Records (EHRs) play an essential role in healthcare systems by centralizing patient information and supporting informed clinical decisions. However, the vast volume and inherent complexity of EHR data create considerable challenges in terms of data interpretation and effective analysis. Recent advancements in Artificial Intelligence (AI), particularly with the development of Large Language Models, offer promising new avenues for researchers to investigate and leverage in this domain [127]. While previous studies have underscored the potential of LLMs in understanding and processing the language of EHRs, a comprehensive scoping review of these capabilities has been notably lacking. To address this issue, a recent study conducts a scoping review through comprehensive analysis, aiming to consolidate the current uses of LLMs and evaluate their significance in EHR-related research [128]. Social determinants of health (SDoH) [129] are also recognized as crucial nonclinical factors that impact patient health risks and clinical outcomes. The incorporation of SDoH data into clinical decision-making holds strong promise for enhancing diagnosis, treatment planning, and overall patient outcomes. However, much of this essential information remains embedded within unstructured clinical notes. Here, NLP has emerged as a primary technology for extracting SDoH data from clinical text, facilitating its integration into patient care practices and broader research. A systematic review provides insight into current NLP approaches and tools that are specifically designed to identify and retrieve SDoH information from unstructured clinical EHR text, emphasizing the vital role of NLP in expanding the practical utility of EHRs [130].

In addition, building a secure and privacy-preserving health data-sharing framework is another critical priority in health care. Successful implementation depends on strict adherence to privacy, confidentiality, and security (PCS) principles, all of which play vital roles. Since PCS terms are often used interchangeably, a recent study clarified their definitions and highlighted their differences. This distinction reinforces the unique importance of each component in managing health data effectively. Leveraging LLM tools for patient consultations could also unlock the potential of free-text information within EHRs, allowing providers to identify safety concerns, support diagnoses, and enhance the quality of clinical-patient interactions. Tailoring LLMs for specific consultation tasks through pre-training and fine-tuning has emerged as a viable and effective approach. One paper discussing this integration suggests that fine-tuned LLMs, combined with effective prompt engineering, could improve the efficiency of clinical consultation processes cost-effectively and strengthen computer-mediated interactions in primary care [131].

Additionally, in the context of discharge summariesŮwhich are often dense and challenging for patients to interpretŮprompt engineering and parameter-efficient fine-tuning of LLMs show great potential in summarizing hospital discharge summaries (HDS) accurately and effectively. This approach, using LLaMA 2 as the base model, enhances the summarization accuracy of medical terms and contexts, generating compact yet comprehensive summaries that could make HDS documentation far more accessible and patient-friendly. The study illustrates how specialized LLMs can refine patient communication through targeted summarization, making it easier for patients to understand their healthcare records [132]. Another key area of research involves the development of effective question-answering (QA) systems for EHRs, which has attracted significant attention in the research community. A methodological review in this domain seeks to identify and analyze existing EHR QA data sets while exploring the primary methodologies used. It also evaluates the metrics applied and highlights the ongoing challenges in EHR QA. This review highlights the progress and unresolved obstacles in the QA field, emphasizing the importance of robust QA models to support EHR interpretation and facilitate clinicians' access to relevant information [133].

Moreover, extracting specific clinical parameters, such as headache frequency, from EHRs presents an ongoing challenge due to variability in clinical documentation. For instance, the accurate extraction of headache frequency a key measure for evaluating responses to migraine treatment is complicated by inconsistent documentation practices. Traditional NLP algorithms struggle to consistently capture these details, underscoring a need for enhanced methodologies tailored specifically to EHR complexities [134]. Another recent study explored the use of zero-shot prompt engineering combined with retrieval-augmented generation (RAG) in generative AI models. This approach was designed to streamline tasks such as summarizing both structured and unstructured data within EHRs and efficiently extracting important malnutrition information, offering promising solutions for reducing manual data processing [135, 136].

To summarize, these studies collectively highlight the evolving role of LLMs and NLP in tackling the complex challenges associated with Electronic Health Records (EHRs). From extracting social determinants of health to enhancing summarization, question-answering, and various data extraction tasks, the advancements in LLM and NLP technologies have demonstrated their potential to transform EHR data management. Continued innovation in this field is essential, as it promises to further improve the accessibility, accuracy, and usability of EHR data, ultimately benefiting both clinicians and patients by streamlining workflows, supporting informed decision-making, and improving healthcare outcomes.

5.2.3 | Medical Literature Synthesis

Medical literature synthesis plays a crucial role in distilling vast amounts of research into concise, actionable insights for healthcare professionals, supporting evidence-based practice, and improving decision-making. With the rapid expansion of medical knowledge, advanced NLP and large language models are increasingly employed to automate the process, making it more efficient and scalable. One such effort is explored in [137], where researchers used ChatGPT 4.0 to generate a comprehensive review of wearable devices for arrhythmia management. The study aimed to understand the feasibility of using LLMs for summarizing academic medicine fields. Simple prompts guided the LLM in producing the review, which was then evaluated against a human-authored counterpart. Independent readers assessed the output, and findings revealed that LLMs could pragmatically support scholarly work when guided by expert inputs, suggesting promising applications in medical literature synthesis.

Expanding on the capabilities of LLMs in medical literature synthesis, researchers in [138] developed GatorTronGPT, a generative clinical LLM, trained on a vast data set of 277B words. This included 82B words of clinical text from over 2M patients at the University of Florida Health, as well as 195B words from diverse general English sources. Built on GPT-3 architecture with 20B parameters, GatorTronGPT was designed to handle biomedical natural language processing (NLP) tasks and healthcare text generation. The system was evaluated for its effectiveness in medical text synthesis and showcased the potential of specialized clinical LLMs to enhance the efficiency and accuracy of medical literature synthesis. Building on these advancements, the study in [139] explores how LLMs can streamline the scientific review process by automating various stages. Researchers conducted a systematic review of projects focused on automating literature review tasks using LLMs, highlighting their potential to significantly reduce the time and effort involved in processing large bodies of medical research. Their findings were encouraging, as the automation of review processes showed promise for improving the speed and accuracy of literature synthesis. As these tools continue to evolve, they are expected to transform how scientific reviews are conducted, allowing for faster and more efficient synthesis of medical knowledge.

The study in [140] assessed the performance of OpenAI's GPT and GPT-4 APIs in identifying relevant titles and abstracts from clinical review data sets. The researchers compared the LLM's results with ground truth labeling performed by two independent human reviewers. The LLMs demonstrated high accuracy, a moderate macro F1-score, and strong sensitivity for excluding irrelevant papers and including relevant ones. These results underscore the effectiveness of LLMs in assisting medical literature synthesis, particularly in tasks such as identifying relevant studies for clinical reviews, further supporting their role in automating evidence-based research processes. The studies reveal that LLMs can play a valuable role in medical literature synthesis when properly guided. From generating reviews and handling clinical text to automating scientific review processes, LLMs offer significant potential to streamline and enhance the synthesis of medical knowledge, improving accessibility for healthcare professionals and researchers.

5.3 | Patient Interaction and Support

The integration of Large Language Models and Natural Language Processing in health care significantly enhances patient interaction and support [141]. By facilitating efficient communication between patients and LLM-based systems, these technologies streamline the process of obtaining medical advice and enhance patient engagement in general as shown in Figure 13 Through individualized, real-time interactions, LLMs and NLP, are revolutionizing patient care by making health care more adaptable and accessible [142]. One significant benefit is the ability to respond instantly to frequently asked questions, allowing patients to get reliable medical advice without wasting time waiting for a medical expert. This enables individuals to autonomously manage their health issues by providing information on symptoms, medications, treatments, and general medical advice [143].

In order to encourage healthy choices and avert future health problems, LLM-powered systems also provide tailored wellness recommendations for programs based on individual medical records and lifestyles. LLM technology effectively manages claims for insurance, billing, and queries on the administrative front, reducing the workload for both patients and physicians. Automated systems help patients submit claims, comprehend coverage, and answer issues about payments [144]. Additionally, by identifying open times, making appointments, and sending reminders, these virtual assistants improve accuracy and ease in appointment scheduling. LLM and NLP systems are excellent at collecting data about patients before consultations, guaranteeing that medical professionals have all the information they need

FIGURE 13 | This diagram outlines the architecture of a medical chatbot system leveraging LLMs for patient interaction, highlighting the integration of NLP modules with real-time medical data access to support clinical communication.

beforehand. These systems collect patient input following visits to enhance services [90]. LLMs' scalability allows for 24/7 assistance, giving patients continuous access to care and increasing the flexibility of services. Patient's relationships with their health are strengthened through personalized service, which includes customized follow-up treatment, prescription reminders, and advice on health management. Modern healthcare delivery is being transformed by LLMs and NLPs, who are enhancing access to real-time information and emphasizing patient-centric approaches.

5.3.1 | Virtual Health Assistants

Virtual health assistants powered by LLMs offer an efficient and user-friendly way for patients to engage in natural language conversations with AI-based healthcare systems [145]. These virtual assistants (Figure 14) use Natural Language Processing (NLP) to understand the nuances of a patient's natural language, which helps to make conversations more natural [146]. For example, virtual assistants can analyze inputs from patients who explain their symptoms in everyday language to offer pertinent health information. In addition to managing appointment scheduling, the chatbots can respond to questions about health, provide information about symptoms, and provide tailored medical advice [147]. Large medical databases and sophisticated algorithms are used by the underlying AI to assess patient input and offer tailored recommendations, improving patient convenience and accessibility to health care [85]. These helpers provide people the confidence to take charge of their health and provide vital resources for illness prevention and control.

Several advanced models enhance the functionality of virtual health assistants. Providing real-time symptom analysis and diagnostic recommendations, Med|Primary AI Assistant [148] distinguishes itself and guarantees that patients receive prompt attention to their medical issues. However, complicated medical problems may be difficult for these tools to handle, which could result in erroneous advice. This emphasizes the need for continuous improvement and clinical supervision to increase the tool's dependability. Ada Health [149] additionally employs a conversational interface to assist patients in evaluating their symptoms to help them narrow down potential diagnoses. Another noteworthy example is Babylon Health [150], which tracks chronic diseases over time and provides patients with individualized guidance and long-term support for treating persistent health difficulties. A more recent development in this domain is the Healthcare Copilot [151], which deploys general-purpose LLMs for real-time medical consultations, showcasing the potential of LLM-based agents to act as interactive and adaptive healthcare assistants.

A virtual avatar called Molly [152], developed by Sensely, is another noteworthy example of this. This customized approach is especially beneficial for patients with chronic conditions because Molly provides ongoing monitoring and direction. Furthermore, Buoy Health [153] provides individualized health advice and diagnostic recommendations by analyzing patient input in real time. Each of these tools demonstrates how virtual assistants can adapt to individual patient needs, improving patient engagement and healthcare accessibility. However, the accuracy of AI-generated advice and data privacy are significant issues raised by the usage of virtual health assistants. Strong security measures are necessary to prevent data misuse in these systems, which handle sensitive personal health data. Additionally, AI-generated medical advice may not always be accurate, particularly for rare conditions. For this reason, these systems need to be continuously enhanced to guarantee reliability and safety.

5.3.2 | Symptom Checkers

Symptom checkers are a vital application of Large Language Models in health care which enables patients to input their

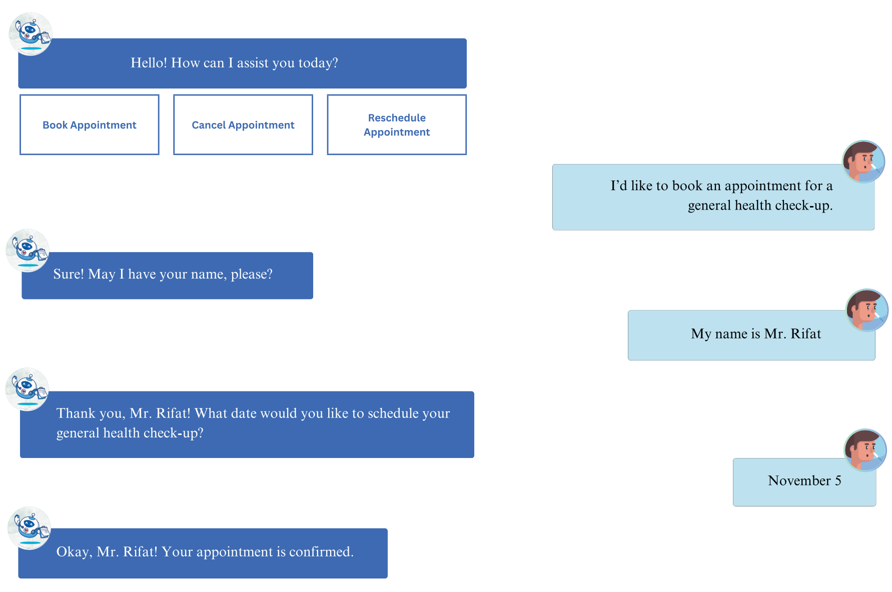

FIGURE 14 | The illustration shows a conversational flow where a patient books a health check-up using a chatbot, highlighting the chatbot's ability to understand natural language and facilitate appointment scheduling.

symptoms and receive immediate feedback on potential diagnoses. These tools primarily function through natural language processing, allowing users to describe their symptoms in everyday terms. They also support structured input formats, such as checkboxes, drop-down menus, images, video, and audio, making them accessible to a broader range of patients. The development of models like GPT-4 [154], GPT-3.5 [155], and Gemini [156] has significantly enhanced the capabilities of symptom checkers. These LLMs excel at interpreting patient-reported symptoms and generating potential diagnoses for common conditions. Research indicates that GPT-4 achieves high diagnostic accuracy due to its training on extensive and diverse medical data sets, enabling effective correlation between symptoms and health conditions. Similarly, Gemini demonstrates precision in disease triage, furthering its reliability for patients seeking timely health guidance [157].

Prominent LLM-based symptom checkers, such as MedPrompt and MedPaLM-2, provide real-time personalized feedback based on user inputs [158, 159]. By simplifying complex medical language into accessible information, these tools help patients better understand their health conditions and make informed decisions. They are also integrated into clinical environments, assisting in data gathering from patient interactions and electronic medical records to create comprehensive diagnostic profiles [160]. Despite these advancements, symptom checkers face challenges, particularly with diagnosing complex conditions. Ongoing research focuses on developing methodologies for evaluating LLM performance in real-world clinical contexts to ensure they meet healthcare standards. Data privacy raises serious ethical issues thus strong protections are necessary to stop abuse. Additionally, managing potential biases in LLM-generated responses is critical for ensuring equitable healthcare delivery. Addressing these challenges is vital for building patient trust and ensuring that symptom checkers provide reliable, high-quality care.

5.3.3 | Natural Language Understanding in Healthcare Chatbots

The integration of Natural Language Understanding (NLU) in healthcare chatbots marks a transformative phase in patient interaction and service delivery. Modern NLU-equipped chatbots, in contrast to simple automated systems, make use of sophisticated deep-learning models to efficiently comprehend language and context by enabling meaningful, patient-centered conversations. With this fundamental change, chatbots may now interact with users in a context-aware manner that seems more natural and human-like, surpassing pre-programmed responses [161]. Central to this evolution is deep-learning models such as BERT [162] and its domain-specific adaptations like SciBERT [163]. These models are made to process and understand complicated medical language to provide precise intent identification and entity extraction. Healthcare-specific NLU architectures (Figure 15) perform better than general NLP systems because they can understand the complex language of medical inquiries and provide accurate answers. This capability is crucial, as correctly interpreting user input impacts the chatbot's reliability and user trust.

Strong NLU integration has important practical ramifications for patient accessibility. Advanced NLU-enabled chatbots can reduce needless trips to medical institutions for non-urgent issues by giving patients timely, evidence-based advice. In addition to saving patients time, this strategy relieves medical staff of some of their workload so they may concentrate on more urgent cases. Using chatbots to automate common questions promotes a smooth information flow, assisting patients in making informed decisions and exercising their autonomy [164]. Beyond interacting directly with patients, NLU also helps healthcare providers. Chatbots can act as a first point of contact for basic information requests, freeing up medical staff to focus on more difficult,

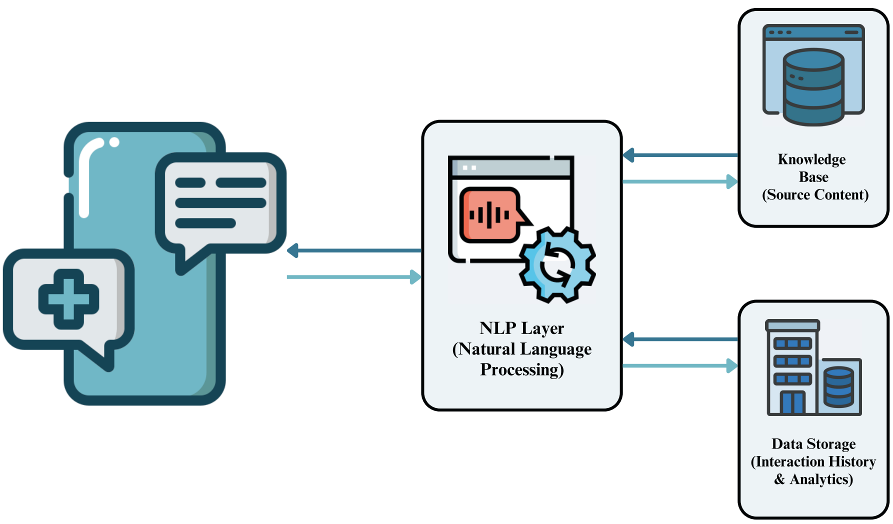

FIGURE 15 | The figure shows a chatbot system that uses an NLP layer to interpret patient inputs, drawing on a knowledge base and interaction history. This setup enables context-aware, personalized responses to support clinical decision-making and patient communication.

urgent duties. This feature helps healthcare systems manage workflow better, which boosts productivity [165]. The potential of NLU in healthcare chatbots continues to grow with emerging frameworks and technologies. The incorporation of specialized models and continuous improvements in NLU frameworks open the door to chatbots that can manage intricate medical situations and offer assistance that closely reflects human knowledge. This technical advancement represents a change toward a more patient-centered strategy in which technology connects the dots between user accessibility and healthcare services.

5.4 | Drug Discovery and Pharmacology

Drug discovery and pharmacology are pivotal fields in health care, where advanced data analysis and interpretation can drive substantial improvements. The integration of large language models (LLMs) and natural language processing (NLP) techniques into these areas offers promising enhancements, particularly for tasks involving the processing of complex biomedical data. LLMs are increasingly leveraged to support the development of pharmacological agents by serving as extensive databases of biological and chemical information. In drug discovery, this technology facilitates the identification of potential drug candidates through predictive modeling and synthesis generation. Drug discovery itself is an interdisciplinary process that encompasses pharmacology, chemistry, biology, and computational science, aiming to identify novel therapeutic compounds [166]. Specialized LLMs have become indispensable for accelerating various chemical assessments essential to drug development [167]. Pharmacology, the scientific discipline focused on understanding how drugs and chemicals interact with biological systems, plays a critical role in this context. Here, a "drug" refers to any chemical substance-natural or synthetic that exerts an effect on a biological system Here, a "drug" refers to any chemical substanceŮnatural or syntheticŮthat exerts an effect on a biological system [168].

5.4.1 | Text Mining for Drug Interactions

Text mining leverages advanced computational methods to analyze vast collections of textual data, identifying key concepts, emerging trends, and previously hidden relationships within biomedical literature and electronic health records (EHRs) [169]. In pharmacology, text mining powered by Large Language Models has become essential in identifying potential drug interactions, especially in large-scale data sets where conventional pharmacovigilance approaches may be limited by labor-intensive data review and manual curation [170]. By processing vast amounts of textual information, LLMs uncover complex patterns and associations between drugs with a speed and accuracy unmatched by traditional techniques.

Several LLMs, such as BioBERT and MedPaLM, have been fine-tuned for biomedical applications, enhancing their role in pharmacovigilance by enabling high-precision analysis of complex drug interactions. Techniques like entity recognition and relationship extraction empower these models to identify drug-related entities, categorize them, and map out potential interactions based on a data set's contextual framework [171]. Furthermore, these models can swiftly identify patterns across millions of text entries, uncovering both established and novel drug-drug interactions (DDIs) more efficiently than traditional pharmacovigilance. The analysis of drug interactions through text mining utilizes computational techniques and tools to examine linguistic patterns in information retrieval derived from annotated corpora [172]. These approaches are particularly effective in flagging DDIs, which may result from two drugs influencing the same genetic pathway or metabolic enzyme [173]. Such capabilities of LLMs in text mining play a pivotal role in preempting adverse drug events, ultimately contributing to improved clinical outcomes and safer pharmacological practices.

5.4.2 | Drug Discovery

Large Language Models are revolutionizing the drug discovery process by offering sophisticated tools for data analysis, target identification, compound screening, and drug design [174]. One of the main contributions of LLMs lies in their ability to integrate diverse data setsŮranging from genetic databases to chemical librariesŮallowing researchers to predict which compounds may effectively interact with biological targets [175]. Virtual screening is another key application of LLMs in drug discovery by predicting molecular interactions and suggesting compounds with high efficacy for specific biological targets. LLMs reduce the need for costly, labor-intensive physical testing of compounds [176]. For instance, through structure-activity relationship (SAR) [177] modeling, LLMs can provide insights into how modifications in a compound's structure might enhance efficacy, lower toxicity, or reduce side effects. This SAR-driven guidance helps optimize drug candidates before they enter the more rigorous stages of development, improving both their effectiveness and safety. The use of LLMs in drug discovery has a transformative impact on the timeline and cost required to bring new treatments to market. By expediting the drug discovery cycle from target identification to compound refinement, LLMs can significantly shorten development timelines, ultimately accelerating the availability of life-saving treatments. This integration of LLMs into the drug discovery pipeline not only reduces costs associated with lengthy clinical trials but also fosters a more agile approach to addressing emerging health threats and unmet medical needs.

5.5 | Diseases Detection

Large Language Models in combination with Natural Language Processing (NLP) techniques are redefining the landscape of medical diagnostics [178], providing transformative solutions across various domains. These advancements have significantly improved the precision, efficiency, and interpretability of disease identification, facilitating timely interventions and enhancing patient care outcomes. Through the integration of both structured and unstructured medical data, LLMs are enabling nuanced pattern recognition that supports early disease detection, differential diagnosis, and actionable insights for preventive care [179]. One critical application of LLMs is in the field of medical imaging, where these models enhance the interpretative accuracy of radiology [180], pathology, and other diagnostic imaging modalities. Figure 16 shows the framework of the detection of diseases using LLM. It offers rapid disease detection that is invaluable in complex cases by cross-referencing image data with extensive medical databases.

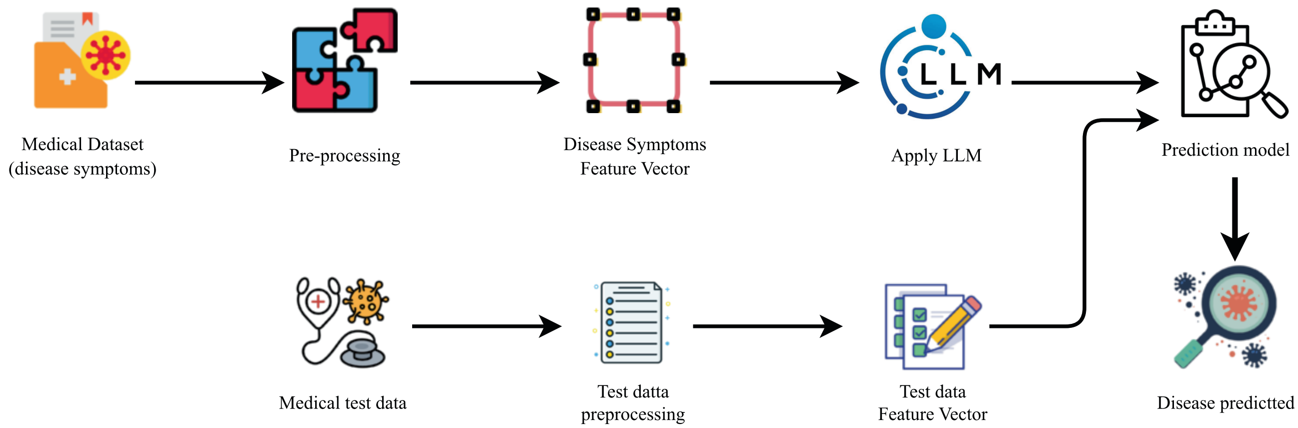

FIGURE 16 | This diagram illustrates how symptom and test data are preprocessed into feature vectors, then analyzed by an LLM and prediction model to identify diseases, highlighting the role of LLMs in automated diagnostics.

For instance, Alzheimer's disease diagnostics can benefit from an LLM framework that extracts and interprets linguistic markers [181, 182]. Such a framework allows clinicians to monitor linguistic shifts and other cognitive markers over time. Additionally, LLMs extend beyond traditional medical settings to fields such as agricultural disease management integrating with object detection and real-time monitoring systems. These models contribute to proactive disease management by analyzing crop health data which supports remediation strategies that prevent widespread crop damage [183, 184]. In this context, LLMs can process large data sets from sensors and image data to identify disease indicators in plants, offering farmers timely and location-specific insights. The application of LLMs in disease detection is not without challenges, processing multimodal data across vast patient data sets presents a significant barrier to real-time clinical implementation.

5.5.1 | Medical Imaging

Medical imaging is a method and procedure that involves visualizing the function of certain organs or tissues as well as imaging the inside of a person for clinical analysis and medical intervention. It is an image retrieval technique that helps in medical disease detection and diagnosis. The use of LLMs in medical imaging [185] is significantly enhancing diagnostic accuracy by automating and optimizing the analysis of complex visual data. Figure 17 illustrates that medical image reports also can be generated using LLMs [186]. Through deep learning and NLP techniques, LLMs can interpret radiology reports, pathology slides, and other forms of medical imaging, providing a second layer of analysis that assists clinicians in making informed decisions. This integration of image processing and language models streamlines the diagnostic workflow, reducing the risk of human error and enabling early disease detection.

i *Automated Image Annotation and Segmentation*: LLMs, when combined with image processing architectures like U-Net or convolutional neural networks (CNNs), can automate the annotation and segmentation of medical images. For example, in oncology, these models can highlight tumor boundaries in radiology scans, distinguishing between malignant and benign regions based on previously annotated data [187]. This capability not only speeds up the diagnostic process but also ensures a high level of consistency in identifying key image features.

ii *Disease Detection and Prognosis Models*: In clinical settings, LLMs aid in detecting specific diseases such as pneumonia, cardiovascular conditions, and neurological disorders by analyzing patterns in imaging data. For example, automated detection of lung nodules in CT scans [188] or signs of Alzheimer's in MRI images allows for early intervention [189]. This predictive capability is particularly useful in screening programs, where quick and accurate assessment of large patient volumes is essential. By enhancing the ability to interpret both visual and textual data, LLMs integrated with medical imaging technologies are driving advancements in diagnostic precision. Models such as U-Net coupled with LLMs enable accurate segmentation of images [190], guiding clinicians in identifying abnormalities that may not be visible to the naked eye. This AI-driven approach provides timely support in areas such as oncology and neurology, where early detection is crucial for successful treatment outcomes.

5.5.2 | Multimodal Data Integration

LLMs are increasingly being employed in multimodal data integration, which combines multiple data types such as textual information, visual data, and audio data from medical imaging to provide a comprehensive perspective on patient health [191]. Figure 18 depicts that the integration allows for a richer diagnostic context, offering a more refined view of a patient's condition than could be achieved from either data type alone. By processing diverse data formats, LLMs enhance diagnostic capabilities, particularly in complex cases that require synthesizing multiple forms of medical evidence. Multimodal LLMs (MLLMs) [192] combine structured data and unstructured data with imaging data. For example, a multimodal model may analyze an MRI scan alongside a patient's history of symptoms and lab results, identifying patterns that suggest a specific diagnosis [193]. This fusion provides deeper insights, especially for diseases that manifest through both visible symptoms in imaging and complex biochemical markers. Multimodal LLMs are particularly useful in diagnosing complex diseases like cancer, cardiovascular conditions, and rare genetic disorders. By integrating data from imaging, lab tests, and patient records, these models can detect disease indicators that may otherwise be overlooked.

For instance, in oncology, LLMs can analyze tumor characteristics from imaging data while considering genetic information and clinical notes, leading to a more precise assessment of cancer stages and treatment options. LLMs such as CLIP (Contrastive Language-Image Pretraining) [194] and Vision-Language Transformers [195] have been employed in multimodal frameworks to predict disease progression or assess the risk of complications. These models analyze both visual and textual inputs to generate predictive insights, such as estimating the likelihood of disease progression in patients with chronic conditions. This predictive power supports preventive care by identifying high-risk individuals based on comprehensive data. Multimodal LLMs represent a significant advancement in disease detection, blending image analysis with textual data to improve diagnostic accuracy.

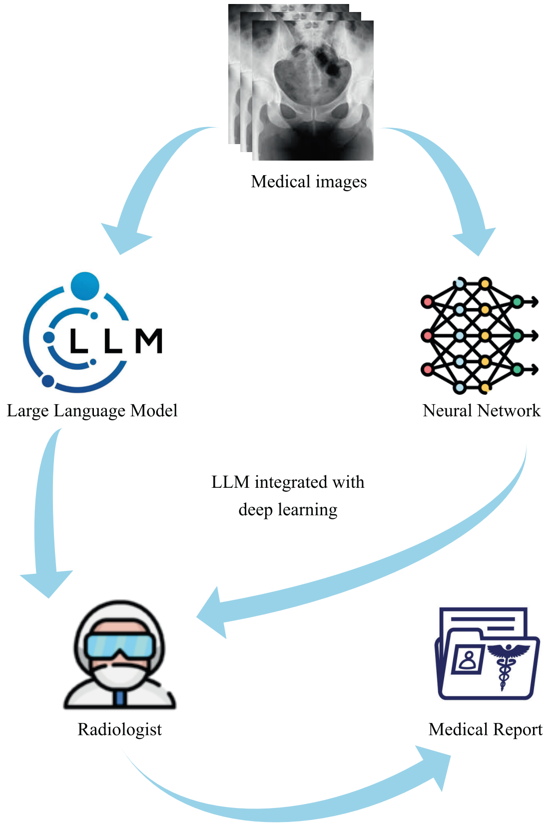

FIGURE 17 | This diagram shows how medical images are processed through neural networks and LLMs to assist radiologists in generating accurate and structured medical reports, enhancing diagnostic efficiency.

5.6 | Public Health and Epidemiology

Epidemiology is a scientific approach focused on the distribution and causes of health and illness within a population [196]. It also satisfies a significant societal need because it identifies disease causes and provides strategies for defending, minimizing, managing, and treating illnesses [197]. In the modern world, epidemiology is no longer the study of disease detection; it is also the science of predicting the occurrence of a crisis, evaluating the risks to the health of the populace, and improving health care measures. Large Language Models (LLMs) and Natural Language Processing (NLP) are combining to improve several facets of health care, thereby changing epidemiology and public health [198]. LLMs can process patient information, enabling more accurate symptom assessment, diagnosis, and personalized treatment planning as shown in Figure 19. They improve patient consultations by automating data retrieval and assisting in clinical trial matching, ensuring that patients receive the most relevant care options [141]. LLMs and NLP also play a critical role in monitoring virus outbreaks by analyzing vast data sources in real-time, contributing to risk analysis and early detection of potential health crises [199]. Additionally, they support health education and training by synthesizing medical knowledge for professionals, while streamlining clinical documentation and medical research tasks. Through these applications, LLMs and NLP enhance both public health preparedness and healthcare delivery, ensuring more efficient and informed decision-making across the healthcare spectrum.

5.6.1 | Surveillance of Emerging Diseases

Surveillance of Emerging Diseases works with language models (LMs) and NLP to monitor, identify, and forecast reactive

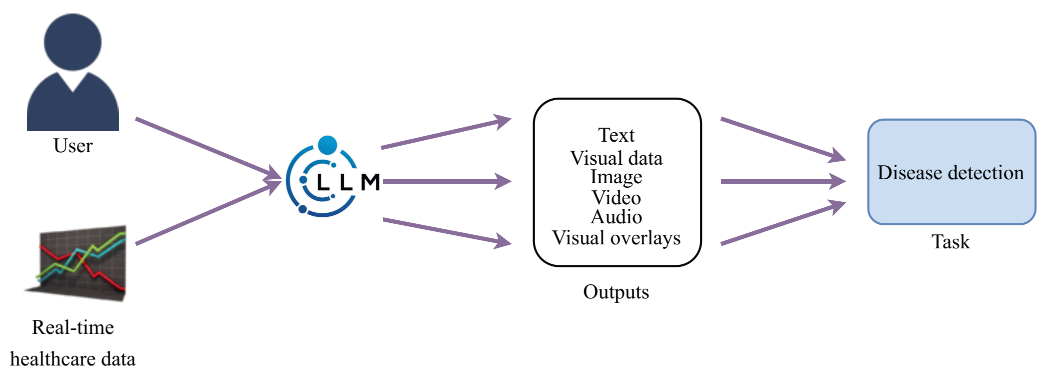

FIGURE 18 | This figure illustrates how LLMs process inputs from users and real-time healthcare data to generate diverse outputs (e.g., text, images, video), enabling downstream tasks such as disease detection and clinical decision-making.

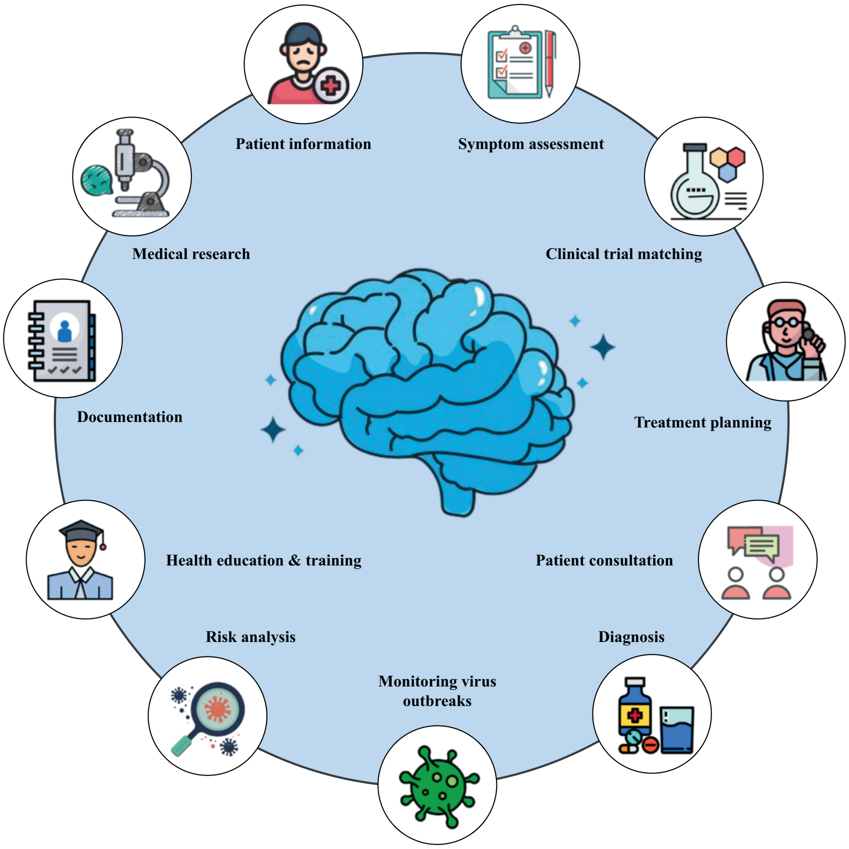

FIGURE 19 | This visual illustrates the wide-ranging uses of LLMs in healthcare, including clinical trial matching, diagnosis, documentation, patient consultation, symptom assessment, and medical research. These applications highlight the transformative potential of LLMs in improving clinical workflows, supporting healthcare decision-making, and enhancing personalized care.

diseases [200]. Real-time information on symptoms and public health is obtained from several sources: clinical reports, social media, and wearable devices. These sets of data are preprocessed where data cleaning and data normalization take place [201]. The next step involves applying various natural language processing techniques, like entity recognition [202], topic modeling, and sentiment analysis, to extract some important features, diseases mentioned [203], their symptoms, or even people's feelings about particular health risks [204]. This aids in identifying recurrences that may eventually reveal one or more possible outbreaks. These will be followed by predictive modeling, where disease spread and risks to various groups will be predicted. Such predictions enable the public health officers to take preventive measures such as giving early warnings and developing communication strategies and policy measures, respectively, as visualized in Figure 20. Continuous learning is an essential part of this system [205]. As new data becomes available, the forecasts change, developing predictions that might not be achievable otherwise. However, the process ran into certain data privacy difficulties to guarantee that personal health data will be treated appropriately and morally. Proper regulations regarding privacy and data anonymization techniques [206] are needed to stop the misuse or illegal access to private health information.

5.6.2 | Predictive Analytics in Epidemiology

Predictive analytics in epidemiology using Large Language Models is transforming disease detection, management, and prevention. These models utilize advanced natural language processing (NLP) techniques. They analyze large amounts of health-related data, including patient symptoms, medical histories, lab results, and environmental factors [207]. This helps in the early identification of potential outbreaks and individual-level disease prediction. LLMs process both structured and unstructured data. This data can come from clinical reports, social media posts, and wearable devices. LLMs detect trends, correlations, and hidden patterns that could signal a health crisis [208]. Real-time data like symptoms and medical records are inputted into LLMs. The models then analyze this information to predict diseases such as lung cancer, heart disease, viral infections, and glaucoma as depicted in Figure 21. The output helps healthcare professionals diagnose conditions early. It offers personalized healthcare solutions and improves decision-making efficiency. By leveraging LLMs, the healthcare system can detect diseases sooner [90], provide tailored treatments, and optimize public health responses, making predictive analytics a vital tool in modern epidemiology [209]. By examining the patient's health data alongside wider

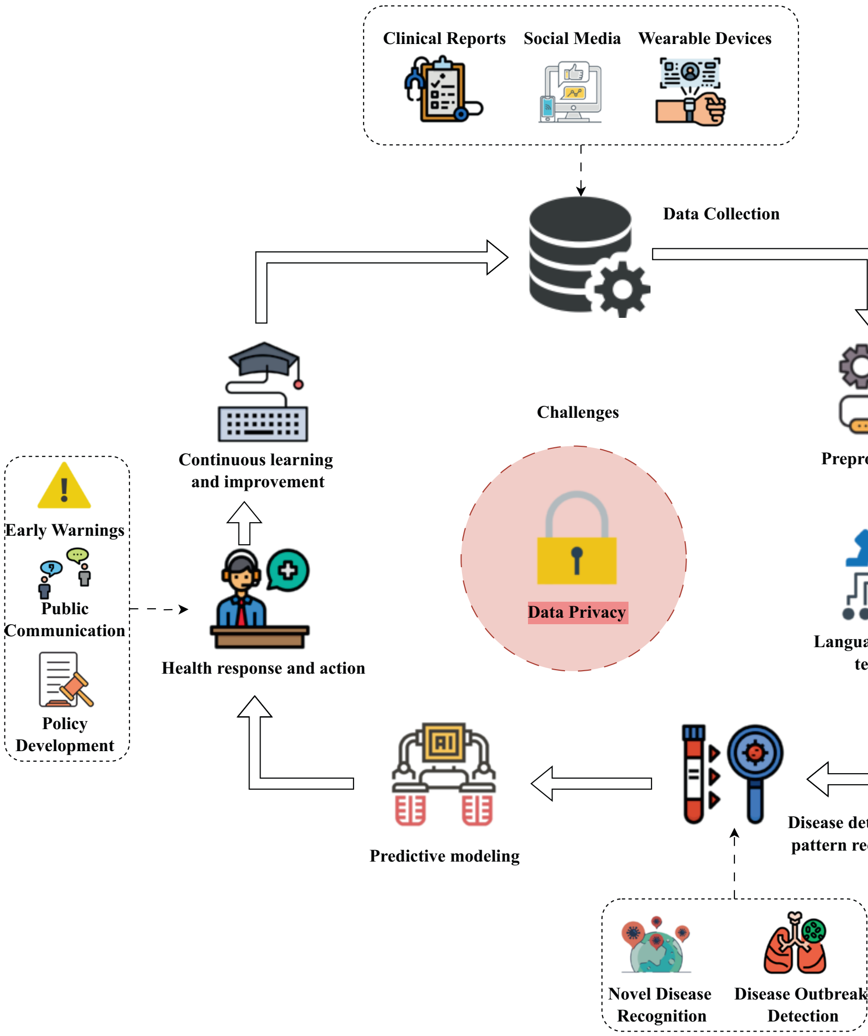

FIGURE 20 | The highlight of how language models and NLP can support early illness surveillance, predictive analysis, and intervention tactics, offering vital data-driven insights for managing and averting the emergence of new diseases.

epidemiological trends, LLMs help doctors make more informed decisions regarding treatment options, medication adjustments, and preventative measures [210]. This capability enhances effective medical approaches, where action can be tailored to each patient's specific needs.

6 | Challenges to Implementation of Generative AI LLM

Generative AI models, Large Language Models in particular have the potential to revolutionize health care by improving patient care, clinical decision-making, and administrative procedures [211]. Still, there are significant obstacles to their incorporation into practical healthcare applications that must be overcome to guarantee their safe, efficient, and moral use. Important issues in this area are data privacy, interpretability, and deployment complexity, all of which are necessary for successfully and responsibly optimizing the advantages of medical LLMs [212].

6.1 | Data Privacy

Data privacy stands as one of the most critical challenges in deploying LLMs within health care. Healthcare data is sensitive by nature since it contains personal health information (PHI), which needs to be protected in compliance with stringent regulations like GDPR in Europe and HIPAA in the US [213]. Large volumes of data are needed to train and optimize generative AI models, and if this data is handled carelessly, patients may be in danger of data breaches, misuse, or illegal access. Since LLMs are trained on large datasets, there is a significant chance that they would unintentionally keep sensitive material, which is especially problematic in the healthcare industry where patient trust is crucial. Solutions to data privacy concerns using federated learning [214], restricting access to models through secure, encrypted environments, and putting strong data anonymization techniques into practice are all part of LLM deployment. Federated learning keeps patient data local and eliminates the need for extensive data transfers by enabling AI models to be trained on

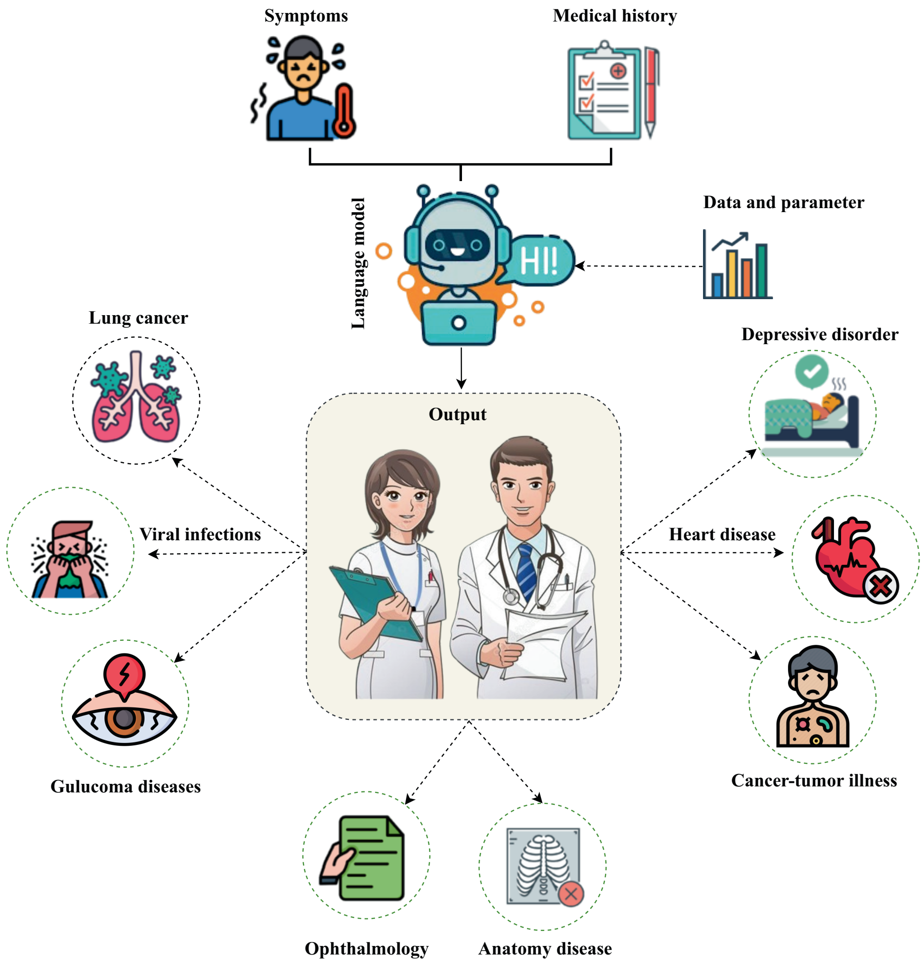

FIGURE 21 | This figure illustrates how large language models, integrated with predictive analytics, process symptoms, medical history, and clinical parameters to assist healthcare professionals in predicting various diseases—ranging from heart disease and cancer to glaucoma and viral infections—enhancing early detection and personalized treatment planning.

decentralized data sources. This method protects patient privacy while allowing LLMs to learn from a variety of data sources without allowing the data to leave safe healthcare settings. To improve accountability and preserve patient trust, procedures for regular audits and open management of AI-generated data must also be put in place [215].

6.2 | Interpretability of LLM

Interpretability is another key barrier to applying LLMs within healthcare settings. The "black boxes" [216] nature of LLMs makes it hard for doctors and other medical experts to comprehend the logic underlying their answers. This is in contrast to traditional medical decision-making procedures, which primarily rely on open and validated processes. Lack of interpretability could result in poorly informed or unduly dependent clinical practices, which could negatively impact patient outcomes in clinical settings where decisions frequently have life-altering consequences. The lack of an easily traceable reasoning process makes verifying and validating AI-provided insights more difficult when LLMs make mistakes or fabricate answers. One way to address interpretability is to develop explainability frameworks to help users comprehend LLMs' decision-making procedures, especially in complex situations. The "Chain of Thought" (CoT) [217] prompting method is one strategy that encourages models to deconstruct their answers into sequential reasoning processes so that medical professionals can assess the model's rationale before acting. Furthermore, studies on uncertainty-aware LLMs seek to show how confident or doubtful these models are in their answers. To help physicians evaluate the AI's recommendations and consider them in the context of a larger clinical setting, this can be combined with tools that emphasize important input parameters for a particular response [218].

6.3 | Deployment of LLM

Deploying medical LLMs environments involves technical, operational, and ethical complexities that require careful consideration. Since the healthcare industry is always changing, keeping the model current with the most recent medical recommendations, therapies, and protocols is one of the main technical problems. Due to their training cutoffs and lack of real-time data access, several LLMs, including ChatGPT, may produce inaccurate information that puts patient safety at risk as well as the standard of treatment. Furthermore, healthcare organizations with limited funding or technological assistance may find it difficult to integrate these models into their current clinical workflows and health information systems due to the large infrastructure modification and resource commitment required [219]. Strong plans for regular model upgrades and specialized personnel are required to address deployment issues and guarantee alignment with current medical knowledge. Although more expensive, implementing on-site medical LLMs systems can reduce dependency on external cloud-based models and provide data security benefits. However, to keep the LLMs up to date with the latest developments in medicine, this method needs to be updated and maintained on a regular basis. To guarantee that AI is utilized as an auxiliary tool rather than as the primary decision-maker, human monitoring should also be a part of AI implementation, especially when it comes to clinical judgments with high risks [199].

6.4 | Clinical Domain-Specific LLM

Large Language Models present tremendous potential for transforming health care but their implementation in clinical domains comes with unique challenges. Medical fields such as ophthalmology, pathology, and personalized treatments require high accuracy, domain-specific knowledge, and strict regulatory adherence, which distinguishes clinical LLMs from general-purpose models. Unlike traditional LLMs, clinical models must be trained on specialized, often limited data sets while also adhering to evolving medical standards. Researchers in [220] explore the application of LLMs in ophthalmology, covering their use in education, research, and clinical practice. The study considers input from various stakeholders, including patients, physicians, and policymakers, and highlights anticipated challenges such as accuracy, biases, interpretability, and data security concerns. As the field evolves, it is critical to define best practice standards collaboratively to ensure patient safety. Building on this, the article in [221] emphasizes the need for distinct regulatory frameworks tailored to medical LLMs. The authors argue that these models, though promising, require a different regulatory approach from traditional AI-based medical technologies. By addressing both the risks and benefits of LLM integration, the article calls for regulations that prioritize patient safety while fostering innovation in medical AI. Ensuring public trust in these systems will be crucial for their widespread adoption and transparent regulations can mitigate risks related to data privacy and unintended harm.

Another study in [222] highlights six critical challenges faced by digital health practitioners when using generative AI (GAI) [223], including issues like bias, hallucinations, privacy concerns, and regulatory compliance. The study goes beyond common topics like adversarial mis-prompting and jailbreaking to address less discussed issues such as over-reliance on generative text models and the ethical implications of using AI-driven systems in health care. Given the sensitive nature of medical data, practitioners must navigate these challenges carefully to avoid adverse outcomes while ensuring the system's reliability. Further insights are provided by a scoping review in [224], which examines the challenges of using LLMs in diagnostic medicine, particularly digital pathology. The review highlights obstacles such as biases in training data, limited contextual understanding, and the interpretability of LLM outputs. These challenges stem from LLMs' inability to fully comprehend medical concepts, leading to issues when deploying them for diagnosis. Additionally, the black-box nature of LLMs limits transparency and hinders the trust of healthcare professionals, exacerbating ethical and regulatory concerns [208]. Ensuring that LLMs are trained on data curated by medical professionals and improving their interpretability is crucial for successful deployment in clinical diagnostics.

In [225], researchers investigate the broader applications of generative AI (GAI) models like ChatGPT and DALL-E in healthcare, identifying real-world use cases such as medical imaging, drug discovery, and personalized treatment plans. While GAI holds promise for various healthcare applications, it faces significant limitations, including inadequate decision-making capabilities, privacy risks, and challenges in integrating with existing healthcare systems. These limitations, coupled with data biases, highlight the need for continued research and refinement of GAI models before they can be reliably deployed in clinical environments. The study also emphasizes future research directions, focusing on improving model accuracy, privacy protection, and integration with healthcare infrastructure. Clinical domain-specific LLMs offer transformative potential in healthcare but they come with significant challenges that need to be addressed. Accuracy, interpretability, biases, and data security remain core concerns, alongside regulatory and ethical considerations. As LLM technologies evolve, collaborative efforts between stakeholders are essential to define standards and develop safe, reliable models that can be integrated into clinical practice while upholding patient safety.

6.5 | Cost and Potential Human Resource

The implementation of Generative AI LLMs in medical and healthcare settings presents considerable financial and operational challenges. A major difficulty is the substantial cost of deploying these models, which requires not only advanced computing resources, such as GPUs or TPUs, but also a robust data management infrastructure to ensure stable and secure model operation. Maintaining these systems demands regular updates, high-capacity storage, and sophisticated management frameworks, all of which add to the overall expense [226]. Equally difficult is the lack of skilled personnel. The effective integration of LLMs into clinical environments necessitates personnel with both advanced AI expertise and a strong background in health care. This dual skill set is essential to accurately interpret LLM outputs and apply them in a clinical context, yet such professionals are currently in limited supply [227]. The need for this specialized workforce introduces a barrier to widespread adoption and places additional training demands on existing healthcare staff. Furthermore, the introduction of LLMs into healthcare workflows raises concerns about potential job displacement, particularly for roles involving routine tasks, such as medical transcription, preliminary report generation, and initial patient interactions. While these models can significantly improve efficiency and reduce workload, they may also lead to a restructuring of traditional job roles. This could create workforce disruptions, necessitating strategic planning, and upskilling initiatives to minimize the risk of job displacement. Balancing these technological advancements with thoughtful workforce management is crucial for harnessing the benefits of AI-driven tools in healthcare without compromising employment stability.

7 | Discussion

Our study underscores the significant promise of Large Language Models in medical and digital health technologies with applications spanning clinical decision support, medical text processing, patient interaction, drug discovery, and disease detection. These models including GPT, BERT, and their variants have already demonstrated their capacity to enhance diagnostic accuracy, reduce administrative workloads, and personalize patient care. However, integrating LLMs into clinical settings poses challenges, particularly concerning data privacy, interpretability, and compliance with healthcare regulations. This gap between promising in silico results and real-world clinical use is often referred to as the AI chasm, which highlights the scarcity of robust clinical validation studies in medical AI [228]. To help bridge this chasm, several practical frameworks have been developedŮsuch as DECIDE-AI (for early-phase evaluation) [229], QAMAI (for multi-dimensional quality assessment) [230], and SMART Prompting (for prompt engineering standards) [231]Ůoffer structured approaches for validating model reliability and deployment readiness. We have incorporated these perspectives to enrich our understanding of real-world barriers.

Our findings align with existing literature which emphasizes the growing role of AI in health care. Prior studies have highlighted the effectiveness of LLMs in processing unstructured data, and our research builds on this by examining specific healthcare applications and their associated challenges. We situate our work within the broader context of digital health transformation, reaffirming the potential of LLMs to improve clinical decision-making and patient outcomes. However, we also stress the need for further refinement of these models to enhance their understanding of medical language and improve interoperability. One of the key limitations of our study is that the LLMs we analyzed are general-purpose models that have not been specifically developed for clinical use. This limitation points to the need for models trained on domain-specific healthcare data which remains scarce and difficult to access. The current models also present issues with reliability and explainability which could impact their real-world deployment in healthcare environments.

We recommend future research focus on the development of multimodal LLMs, integrating text, imaging, and other medical data to improve diagnostic accuracy. Domain-specific models for fields such as oncology, radiology, medical imaging, cardiology, etc. could also prove beneficial and tailored to the unique needs of these areas. Furthermore, human-AI hybrid systemsŮwhere clinicians oversee or interact with LLM outputsŮmay help mitigate hallucinations and boost clinical trust. Additionally, research into the ethical concerns surrounding LLMs, especially regarding data privacy and bias will be crucial in ensuring safe and effective integration into healthcare systems. Improved benchmarking standards and regulatory guidelines specific to medical LLMs are also essential to ensure safe, fair, and reproducible deployment across care settings. While LLMs hold transformative potential for health care, overcoming these challenges will be essential for their successful implementation. Continued collaboration among AI researchers, healthcare professionals, and policymakers is necessary to harness the full benefits of these models while ensuring patient safety and ethical compliance.

8 | Conclusion

In conclusion, this review has provided a thorough examination of the fast-evolving field of large-language medical models, detailing their foundational architectures, recent advancements, and critical considerations for reliability. It has traced the progression of LLMs, highlighted the shift from general-purpose to specialized models, and emphasized their significant potential to revolutionize multiple areas of health care. This review has analyzed key algorithmic advancements, including clinical reasoning, NLP integration, and retrieval-augmented generation, which improve LLMs effectiveness in managing complex medical queries and generating accurate responses. It has also explored the wide-ranging applications of LLMs, from clinical documentation and diagnostic support to patient communication and medical education, highlighting their vast potential to streamline healthcare processes and enhance patient outcomes. Furthermore, the paper has underscored the crucial importance of the trustworthy and safe deployment of medical LLMs, addressing challenges related to fairness, accountability, privacy, and robustness. It has highlighted the need for rigorous evaluation, ethical considerations, and the development of regulatory frameworks. Looking ahead, the review has identified promising research directions, including advancements in algorithms, industry transformations, and policy developments, all aimed at supporting the responsible growth of medical LLMs.

Author Contributions

All authors contributed equally to this manuscript. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors would like to express their sincere gratitude to the members of the AI-Based Medical Imaging Laboratory (AMI Lab) for their continuous support and technical assistance. We also acknowledge the institutional support provided by the Department of Information Convergence Engineering, Pusan National University, Yangsan Campus, ROK. The authors would like to thank the members of the Daffodil International University Multi-Disciplinary Action Lab, as well as the anonymous reviewers, for their valuable feedback on earlier drafts of this article.

Disclosure

All figures in this manuscript are original creations by the authors.

Ethics Statement

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing not applicable to this article as no data sets were generated or analyzed during the current study.

References

1. W. X. Zhao, K. Zhou, J. Li, et al., "A Survey of Large Language Models," 2023. arXiv Preprint arXiv:2303.18223.

2. J. Yang, H. Jin, R. Tang, et al., "Harnessing the Power of Llms in Practice: A Survey on Chatgpt and Beyond," *ACM Transactions on Knowledge Discovery From Data* 18, no. 6 (2024): 1–32.

3. H. Touvron, T. Lavril, G. Izacard, et al., "Llama: Open and Efficient Foundation Language Models," 2023. arXiv Preprint arXiv:2302.13971.

4. J. Achiam, S. Adler, S. Agarwal, et al., "Gpt-4 Technical Report," 2023. arXiv Preprint arXiv:2303.08774.

5. Z. Du, Y. Qian, X. Liu, et al., "Glm: General Language Model Pretraining With Autoregressive Blank Infilling," 94 (2021), https://arxiv.org/abs/2103.10360.

6. K. Singhal, T. Tu, J. Gottweis, et al., "Towards Expert-Level Medical Question Answering With Large Language Models," 2023. arXiv Preprint arXiv:2305.09617.

7. H. Nori, Y. T. Lee, S. Zhang, et al., "Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine," 2023. arXiv Preprint arXiv:2311.16452.

8. D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset From Medical Exams," *Applied Sciences* 11, no. 14 (2021): 6421.

9. S. Wan, H. Lin, W. Gan, J. Chen, and P. S. Yu, "Web3: The Next Internet Revolution," 2023. arXiv Preprint arXiv:2304.06111.

10. J. Sun, W. Gan, H.-C. Chao, S. Y. Philip, and W. Ding, "Internet of Behaviors: A Survey," *IEEE Internet of Things Journal* 10, no. 13 (2023): 11117–11134.

11. C. Zhang, Q. Dai, Z. Du, W. Gan, J. Weng, and S. Y. Philip, "Tusq: Targeted High-Utility Sequence Querying," *IEEE Transactions on Big Data* 9, no. 2 (2022): 512–527.

12. W. Gan, J. C.-W. Lin, H.-C. Chao, and S. Y. Philip, "Discovering High Utility Episodes in Sequences," *IEEE Transactions on Artificial Intelligence* 4, no. 3 (2022): 473–486.

13. M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "The Prisma 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," *BMJ* 372 (2021): n71.

14. A. H. Huang, H. Wang, and Y. Yang, "Finbert: A Large Language Model for Extracting Information From Financial Text," *Contemporary Accounting Research* 40, no. 2 (2023): 806–841.

15. Y. Chang, X. Wang, J. Wang, et al., "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology* 15, no. 3 (2024): 1–45.

16. M. U. Hadi, Q. Al Tashi, A. Shah, et al., "Large Language Models: A Comprehensive Survey of Its Applications, Challenges, Limitations, and Future Prospects," *Authorea Preprints* (2024).

17. L. Yan, L. Sha, L. Zhao, et al., "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review," *British Journal of Educational Technology* 55, no. 1 (2024): 90–112.

18. J. Clusmann, F. R. Kolbinger, H. S. Muti, et al., "The Future Landscape of Large Language Models in Medicine," *Communications Medicine* 3 (2023): 141, https://api.semanticscholar.org/CorpusID:263827375.

19. N. Humza, A. U. Khan, S. Qiu, et al., "A Comprehensive Overview of Large Language Models," 2023. arXiv Preprint arXiv:2307.06435.

20. Z. Zhang, Y. Song, G. Yu, et al., "Relu2 Wins: Discovering Efficient Activation Functions for Sparse Llms," *ArXiv* (2024): abs/2402.03804.

21. V. M. Kushala, H. Warrier, Y. Gupta, et al., "Fine Tuning Llm for Enterprise: Practical Guidelines and Recommendations," 2024. arXiv Preprint arXiv:2404.10779.

22. W. Xu, W. Hu, F. Wu, and S. Sengamedu, "Detime: Diffusion-Enhanced Topic Modeling Using Encoder-Decoder Based Llm," 2023. arXiv Preprint arXiv:2310.15296.

23. H. Jordan, S. Borgeaud, A. Mensch, et al., "Training Compute-Optimal Large Language Models," 2022. arXiv Preprint arXiv:2203.15556.

24. J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoeybi, and S. Han, "Vila: On Pre-Training for Visual Language Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), 26689–26699.

25. M. A. Khan Raiaan, M. S. Hossain Mukta, K. Fatema, et al., "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," *IEEE Access* 12 (2024): 26839–26874.

26. G. Yenduri, M. Ramalingam, G. C. Selvi, et al., "Gpt (Generative Pre-Trained Transformer)–a Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," *IEEE Access* 12 (2024): 54608–54649.

27. S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, and T. Oladunni, "Advancements in Generative Ai: A Comprehensive Review of Gans, Gpt, Autoencoders, Diffusion Model, and Transformers," *IEEE Access* 12 (2024): 69812–69837.

28. R. Luo, L. Sun, Y. Xia, et al., "Biogpt: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining," *Briefings in Bioinformatics* 23, no. 6 (2022): bbac409.

29. Y. Zhang, H. Pei, S. Zhen, Q. Li, and F. Liang, "Chat Generative Pre-Trained Transformer (Chatgpt) Usage in Healthcare," *Gastroenterology & Endoscopy* 1, no. 3 (2023): 139–143.

30. P. Su and K. Vijay-Shanker, "Investigation of Improving the Pre-Training and Fine-Tuning of Bert Model for Biomedical Relation Extraction," *BMC Bioinformatics* 23, no. 1 (2022): 120.

31. Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-Training With Whole Word Masking for Chinese Bert," *IEEE/ACM Transactions on Audio, Speech and Language Processing* 29 (2021): 3504–3514.

32. H. Sebbaq and N.-e. El Faddouli, "Fine-Tuned Bert Model for Large Scale and Cognitive Classification of Moocs," *International Review of Research in Open and Distance Learning* 23, no. 2 (2022): 170–190.

33. G. Jawahar, B. Sagot, and D. Seddah, "What Does Bert Learn About the Structure of Language?," in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2019).

34. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (Curran Associates Inc., 2019), 117.

35. N. Arabadzhieva-Kalcheva and I. Kovachev, "Comparison of Bert and Xlnet Accuracy With Classical Methods and Algorithms in Text Classification," in *2021 International Conference on Biomedical Innovations and Applications (BIA)*, vol. 1 (IEEE, 2022), 74–76.

36. F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based

Emotion Recognition," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (IEEE, 2020), 117–121.

37. M. Lewis, Y. Liu, N. Goyal, et al., "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ed. D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Association for Computational Linguistics, 2020), 7871–7880.

38. D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic Forecasting With Autoregressive Recurrent Networks," *International Journal of Forecasting* 36, no. 3 (2020): 1181–1191.

39. P. P. S. Bedi, M. Bala, and K. Sharma, "Mlm: Masked Language Modeling Using Deep Learning for Efficient Summarization of Unstructured Data," in *International Conference on Data Analytics & Management* (Springer, 2023), 339–347.

40. Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-Xl: Attentive Language Models Beyond a Fixed-Length Context," in *Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2019).

41. K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and Permuted Pre-Training for Language Understanding," in *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates Inc, 2020), 16857–16867.

42. J. Devlin, M.-W. Chang, a. L. K. Kenton, and K. Toutanova, "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of naacL-HLT, 1, 2* (Association for Computational Linguistics, 2019).

43. Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A Robustly Optimized BERT Pre-Training Approach With Post-Training," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, ed. S. Li, M. Sun, Y. Liu, et al. (Chinese Information Processing Society of China, 2021), 1218–1227.

44. A. Radford, J. Wu, R. Child, et al., "Language Models Are Unsupervised Multitask Learners," *OpenAI Blog* 1, no. 8 (2019): 1–24.

45. R. Sennrich, B. Haddow, and A. Birch, "Neural Machine Translation of Rare Words With Subword Units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, ed. K. Erk and N. A. Smith (Association for Computational Linguistics, 2016), 1715–1725.

46. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net* (openreview, 2020).

47. C. Raffel, N. Shazeer, A. Roberts, et al., "Exploring the Limits of Transfer Learning With a Unified Text-To-Text Transformer," *Journal of Machine Learning Research* 21, no. 140 (2020): 1–67.

48. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "Glue: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," 2018. arXiv Preprint arXiv:1804.07461.

49. P. Rajpurkar, "Squad: 100,000+ Questions for Machine Comprehension of Text," 2016. arXiv Preprint arXiv:1606.05250.

50. G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, "Race: Large-Scale Reading Comprehension Dataset From Examinations," 2017. arXiv Preprint arXiv:1704.04683.

51. Z. Lan, M. Chen, S. Goodman, et al., "ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations," 2019, http://arxiv.org/abs/1909.11942.

52. L. Han, X. Zeng, and L. Song, "A Novel Transfer Learning Based on Albert for Malicious Network Traffic Classification," *International Journal of Innovative Computing, Information and Control* 16, no. 6 (2020): 2103–2119.

53. G. Deepanway, M. Navonil, M. Rada, and P. Soujanya, "Stack: Sentence Ordering With Temporal Commonsense Knowledge," 2021, https://arxiv.org/abs/2109.02247.

54. S. Afantenos, V. Karkaletsis, and P. Stamatopoulos, "Summarization From Medical Documents: A Survey," *Artificial Intelligence in Medicine* 33, no. 2 (2005): 157–177, https://doi.org/10.1016/j.artmed.2004.07.017.

55. M. Guevara, S. Chen, S. Thomas, et al., "Large Language Models to Identify Social Determinants of Health in Electronic Health Records," *NPJ Digital Medicine* 7, no. 1 (2024): 6.

56. T. Nakaura, R. Ito, D. Ueda, et al., "The Impact of Large Language Models on Radiology: A Guide for Radiologists on the Latest Innovations in Ai," *Japanese Journal of Radiology* 42 (2024): 1–696.

57. J. Cheng, "Applications of Large Language Models in Pathology," *Bioengineering* 11, no. 4 (2024): 342, https://doi.org/10.3390/bioengineering11040342.

58. M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-Lm: Training Multi-Billion Parameter Language Models Using Model Parallelism," 2019. arXiv Preprint arXiv:1909.08053.

59. D. Narayanan, M. Shoeybi, J. Casper, et al., "Efficient Large-Scale Language Model Training on Gpu Clusters Using Megatron-Lm," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (Association for Computing Machinery, 2021), 1–15.

60. T. B. Brown, "Language Models Are Few-Shot Learners," 2020. arXiv Preprint arXiv:2005.14165.

61. M. Goyal and Q. H. Mahmoud, "A Systematic Review of Synthetic Data Generation Techniques Using Generative Ai," *Electronics* 13, no. 17 (2024): 3509.

62. V. Chamola, G. Bansal, T. K. Das, et al., "Beyond Reality: The Pivotal Role of Generative Ai in the Metaverse," *IEEE Internet of Things Magazine* 7, no. 4 (2024): 126–135.

63. G. M. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A Comprehensive Survey and Analysis of Generative Models in Machine Learning," *Computer Science Review* 38 (2020): 100285.

64. A. Amini, R. Cotterell, J. Hewitt, L. Malagutti, C. Meister, and T. Pimentel, "Generating Text From Language Models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)* (Association for Computational Linguistics, 2023), 27–31.

65. T. Luhman and E. Luhman, "High Fidelity Image Synthesis With Deep Vaes in Latent Space," 2023. arXiv Preprint arXiv:2303.13714.

66. G. Zhong, W. Gao, Y. Liu, Y. Yang, D.-H. Wang, and K. Huang, "Generative Adversarial Networks With Decoder–Encoder Output Noises," *Neural Networks* 127 (2020): 19–28.

67. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2020): 139–144.

68. A. Abu-Srhan, M. A. M. Abushariah, and O. S. Al-Kadi, "The Effect of Loss Function on Conditional Generative Adversarial Networks," *Journal of King Saud University, Computer and Information Sciences* 34, no. 9 (2022): 6977–6988.

69. Y. Liu, G. Qin, X. Huang, J. Wang, and M. Long, "Autotimes: Autoregressive Time Series Forecasters via Large Language Models," 2024. arXiv Preprint arXiv:2402.02370.

70. E. J. Hannan and L. Kavalieris, "Regression, Autoregression Models," *Journal of Time Series Analysis* 7, no. 1 (1986): 27–49.

71. G. Eshel, "The Yule Walker Equations for the ar Coefficients," 2003 Internet resource, 2:68–73.

72. M. Li and X. Liu, "Maximum Likelihood Least Squares Based Iterative Estimation for a Class of Bilinear Systems Using the Data Filtering Technique," *International Journal of Control, Automation and Systems* 18, no. 6 (2020): 1581–1592.

73. Y. Mo, H. Qin, Y. Dong, Z. Zhu, and Z. Li, "Large Language Model (Llm) ai Text Generation Detection Based on Transformer Deep Learning Algorithm," 2024. arXiv Preprint arXiv:2405.06652.

74. Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, and H. Liu, "Natural Language Processing (Nlp) in Management Research: A Literature Review," *Journal of Management Analytics* 7, no. 2 (2020): 139–172.

75. D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language Processing: State of the Art, Current Trends and Challenges," *Multimedia Tools and Applications* 82, no. 3 (2023): 3713–3744.

76. D. Li, N. Yang, W. Wang, et al., "Unified Language Model Pre-Training for Natural Language Understanding and Generation," in *Advances in Neural Information Processing Systems*, vol. 32 (Association for Computational Linguistics, 2019).

77. R. Dale, "Natural Language Generation: The Commercial State of the Art in 2020," *Natural Language Engineering* 26, no. 4 (2020): 481–487.

78. K. R. Chowdhary, "Natural Language Processing," in *Fundamentals of Artificial Intelligence* (Springer, 2020), 603–649. Springer India.

79. Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: Lstm Cells and Network Architectures," *Neural Computation* 31, no. 7 (2019): 1235–1270.

80. H. Zhu, C. He, Y. Fang, and W. Xiao, "Fine Grained Named Entity Recognition via seq2seq Framework," *IEEE Access* 8 (2020): 53953–53961.

81. Y. Wu, J. Xu, M. Jiang, Y. Zhang, and H. Xu, "A Study of Neural Word Embeddings for Named Entity Recognition in Clinical Text," in *AMIA Annual Symposium Proceedings* (American Medical Informatics Association, 2015), 1326.

82. E. Hossain, R. Rana, N. Higgins, et al., "Natural Language Processing in Electronic Health Records in Relation to Healthcare Decision-Making: A Systematic Review," *Computers in Biology and Medicine* 155 (2023): 106649.

83. X. Meng, X. Yan, K. Zhang, et al., "The Application of Large Language Models in Medicine: A Scoping Review," *Iscience* 27, no. 5 (2024): 109713.

84. Q. Chen, J. Du, Y. Hu, et al., "Large Language Models in Biomedical Natural Language Processing: Benchmarks, Baselines, and Recommendations," 2023. arXiv Preprint arXiv:2305.16326.

85. S. AAlowais, S. SAlghamdi, T. A. NadaAlsuhebany, et al., "Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice," *BMC Medical Education* 23, no. 1 (2023): 689.

86. D. Wang and S. Zhang, "Large Language Models in Medical and Healthcare Fields: Applications, Advances, and Challenges," *Artificial Intelligence Review* 57, no. 11 (2024): 299.

87. C. W. Safranek, A. E. Sidamon-Eristoff, A. Gilson, and D. Chartash, "The Role of Large Language Models in Medical Education: Applications and Implications," 2023.

88. N. Kumar, R. K. Kaushal, S. N. Panda, and S. Bhardwaj, "Impact of the Internet of Things and Clinical Decision Support System in Healthcare," in *IoT and WSN Based Smart Cities: A Machine Learning Perspective* (Springer, 2022), 15–26.

89. N. Ahmad, S. Du, F. Ahmed, N. u. Amin, and X. Yi, "Healthcare Professionals Satisfaction and Ai-Based Clinical Decision Support System in Public Sector Hospitals During Health Crises: A Cross-Sectional Study," *Information Technology and Management* 1 (2023): 1–13.

90. P. Kummaragunta Joel, "Integrating Large Language Models for Enhanced Clinical Decision Support Systems in Modern Healthcare," *Journal of Machine Learning for Healthcare Decision Support* 3, no. 1 (2023): 18–62.

91. K. Singhal, S. Azizi, T. Tu, et al., "Large Language Models Encode Clinical Knowledge," 2022. arXiv Preprint arXiv:2212.13138.

92. T. Porat, B. Delaney, and O. Kostopoulou, "The Impact of a Diagnostic Decision Support System on the Consultation: Perceptions of Gps and Patients," *BMC Medical Informatics and Decision Making* 17 (2017): 1–9.

93. S. Reed T, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success," *NPJ Digital Medicine* 3, no. 1 (2020): 17.

94. R. A. Miller, *Diagnostic Decision Support Systems* (Springer International Publishing, 2016), 181–208, https://doi.org/10.1007/978-3-319-31913-1_11.

95. A. S. Rawat, A. Rana, A. Kumar, and A. Bagwari, "Application of Multi Layer Artificial Neural Network in the Diagnosis System: A Systematic Review," *IAES International Journal of Artificial Intelligence* 7, no. 3 (2018): 138.

96. C. Comito, D. Falcone, and A. Forestiero, "Ai-Driven Clinical Decision Support: Enhancing Disease Diagnosis Exploiting Patients Similarity," *IEEE Access* 10 (2022): 6878–6888.

97. B. Zhou, Y. Zhan, Z. Wang, et al., "Benchmarking Medical Llms on Anesthesiology: A Comprehensive Dataset in Chinese," *IEEE Transactions on Emerging Topics in Computational Intelligence* 9 (2025): 3057–3071.

98. F. Bordes, R. Y. Pang, A. Ajay, et al., "An Introduction to Vision-Language Modeling," 2024. arXiv Preprint arXiv:2405.17247.

99. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ed. N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi (Springer International Publishing, 2015), 234–241.

100. C. Wu, Z. Lin, W. Fang, and Y. Huang, "A Medical Diagnostic Assistant Based on Llm," in *Health Information Processing. Evaluation Track Papers*, ed. H. Xu, Q. Chen, H. Lin, et al. (Springer Nature Singapore, 2024), 135–147.

101. I. A. Pap and S. Oniga, "Ehealth Assistant Ai Chatbot Using a Large Language Model to Provide Personalized Answers Through Secure Decentralized Communication," *Sensors* 24, no. 18 (2024): 6140, https://doi.org/10.3390/s24186140.

102. A. Madabhushi and G. Lee, "Image Analysis and Machine Learning in Digital Pathology: Challenges and Opportunities," *Medical Image Analysis* 33 (2016): 170–175.

103. A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. W. L. Aerts, "Artificial Intelligence in Radiology," *Nature Reviews Cancer* 18, no. 8 (2018): 500–510.

104. S. Yu, F. Farooq, A. Van Esbroeck, G. Fung, V. Anand, and B. Krishnapuram, "Predicting Readmission Risk With Institution-Specific Prediction Models," *Artificial Intelligence in Medicine* 65, no. 2 (2015): 89–96.

105. R. Guerra, "Enhancing Risk Management in Hospitals: Leveraging Artificial Intelligence for Improved Outcomes," *Italian Journal of Medicine* 18, no. 2 (2024): 1–12.

106. P. Pan, Y. Wang, C. Liu, et al., "Revisiting the Potential Value of Vital Signs in the Real-Time Prediction of Mortality Risk in Intensive Care Unit Patients," *Journal of Big Data* 11, no. 1 (2024): 53.

107. D. P. Jeong, Z. C. Lipton, and P. Ravikumar, "Llm-Select: Feature Selection With Large Language Models," 2024. arXiv Preprint arXiv:2407.02694.

108. H. Cui, Z. Shen, J. Zhang, et al., "Llms-Based Few-Shot Disease Predictions Using Ehr: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction," 2024. arXiv Preprint arXiv:2403.15464.

109. Y. Dou, Y. Huang, X. Zhao, et al., "Shennongmgs: An Llm-Based Chinese Medication Guidance System," *ACM Transactions on Management Information Systems* 1 (2024): 1–10.

110. Y. Zheng, W. Gan, Z. Chen, Z. Qi, Q. Liang, and P. S. Yu, "Large Language Models for Medicine: A Survey," *International Journal of Machine Learning and Cybernetics* 1 (2024): 1–26.

111. N. Singh, A. Sah, Vartika, et al., "A Hybrid Ayurvedic Drug Recommendation System With Generative Ai," in *International Conference on Innovative Computing and Communication* (Springer, 2024), 175–186.

112. I. Mitra, S. Karmakar, K. Ray, and T. Kar, "Apriori Based Medicine Recommendation System," in *Robotics, Control and Computer Vision: Select Proceedings of ICRCCV 2022* (Springer, 2023), 219–228.

113. R. Ramachandran, V. Chinnammal, K. Malathy, K. Sailaja, and S. Shanmugam, "An Improved Drug Recommendation System Using Artificial Intelligence Assisted Learning Methodology," in *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)* (IEEE, 2024), 1–7.

114. T. I. Wilhelm, J. Roos, and R. Kaczmarczyk, "Large Language Models for Therapy Recommendations Across 3 Clinical Specialties: Comparative Study," *Journal of Medical Internet Research* 25 (2023): e49324.

115. Y. Dou, X. Zhao, H. Zou, J. Xiao, P. Xi, and S. Peng, "Shennonggpt: A Tuning Chinese Llm for Medication Guidance," in *IEEE International Conference on Medical Artificial Intelligence (MedAI)* (IEEE, 2023), 67–72.

116. C. Pais, J. Liu, R. Voigt, V. Gupta, E. Wade, and M. Bayati, "Large Language Models for Preventing Medication Direction Errors in Online Pharmacies," *Nature Medicine* 30 (2024): 1–9.

117. L. Giordano and M. P. Di Buono, "Large Language Models as Drug Information Providers for Patients," in *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024* (Association for Computational Linguistics, 2024), 54–63.

118. A. Biswas and W. Talukdar, "Intelligent Clinical Documentation: Harnessing Generative Ai for Patient-Centric Clinical Note Generation," 2024. arXiv Preprint arXiv:2405.18346.

119. S. Sheikhalishahi, R. Miotto, J. T. Dudley, et al., "Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review," *JMIR Medical Informatics* 7, no. 2 (2019): e12239.

120. F. Pethani and A. G. Dunn, "Natural Language Processing for Clinical Notes in Dentistry: A Systematic Review," *Journal of Biomedical Informatics* 138 (2023): 104282.

121. M. Gholipour, R. Khajouei, P. Amiri, S. H. Gohari, and L. Ahmadian, "Extracting Cancer Concepts From Clinical Notes Using Natural Language Processing: A Systematic Review," *BMC Bioinformatics* 24, no. 1 (2023): 405.

122. T. A. Koleck, N. P. Tatonetti, S. Bakken, et al., "Identifying Symptom Information in Clinical Notes Using Natural Language Processing," *Nursing Research* 70, no. 3 (2021): 173–183.

123. N. Tavabi, J. Pruneski, S. Golchin, et al., "Building Large-Scale Registries From Unstructured Clinical Notes Using a Low-Resource Natural Language Processing Pipeline," *Artificial Intelligence in Medicine* 151 (2024): 102847.

124. S. Prakash Nathaniel Kumar and V. T. Mangam, "Ai-Driven Natural Language Processing in Healthcare: Transforming Patient-Provider Communication," *Indian Journal of Pharmacy Practice* 17, no. 1 (2024): 1–12.

125. N. Tavabi, M. Singh, J. Pruneski, and A. M. Kiapour, "Systematic Evaluation of Common Natural Language Processing Techniques to Codify Clinical Notes," *PLoS One* 19, no. 3 (2024): e0298892.

126. H. S. Suh, J. L. Tully, M. N. Meineke, R. S. Waterman, and R. A. Gabriel, "Identification of Preanesthetic History Elements by a Natural Language Processing Engine," *Anesthesia & Analgesia* 135, no. 6 (2022): 1162–1171.

127. X. Yang, A. Chen, N. PourNejatian, et al., "A Large Language Model for Electronic Health Records," *NPJ Digital Medicine* 5, no. 1 (2022): 194.

128. L. Li, J. Zhou, Z. Gao, et al., "A Scoping Review of Using Large Language Models (Llms) to Investigate Electronic Health Records (Ehrs)," *ArXiv* (2024): abs/2405.03066.

129. P. Braveman and L. Gottlieb, "The Social Determinants of Health: It's Time to Consider the Causes of the Causes," *Public Health Reports* 129, no. 1_suppl2 (2014): 19–31.

130. B. G. Patra, M. M. Sharma, V. Vekaria, et al., "Extracting Social Determinants of Health From Electronic Health Records Using Natural Language Processing: A Systematic Review," *Journal of the American Medical Informatics Association* 28, no. 12 (2021): 2716–2727.

131. P. Kumarapeli, T. Haddad, and S. de Lusignan, "Unlocking the Potential of Free Text in Electronic Health Records With Large Language Models (Llm): Enhancing Patient Safety and Consultation Interactions," in *Digital Health and Informatics Innovations for Sustainable Health Care Systems* (IOS Press, 2024), 746–750.

132. J. Goswami, K. K. Prajapati, A. Saha, and A. K. Saha, "Parameter-Efficient Fine-Tuning Large Language Model Approach for Hospital Discharge Paper Summarization," *Applied Soft Computing* 157 (2024): 111531.

133. J. Bardhan, K. Roberts, and D. Z. Wang, "Question Answering for Electronic Health Records: Scoping Review of Datasets and Models," *Journal of Medical Internet Research* 26 (2024): e53636, https://doi.org/10.2196/53636.

134. C.-C. Chiang, M. Luo, G. Dumkrieger, et al., "A Large Language Model–Based Generative Natural Language Processing Framework Fine-Tuned on Clinical Notes Accurately Extracts Headache Frequency From Electronic Health Records," *Headache* 64, no. 4 (2024): 400–409.

135. P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive Nlp Tasks," in *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates Inc, 2020), 9459–9474.

136. M. Alkhalaf, P. Yu, M. Yin, and C. Deng, "Applying Generative Ai With Retrieval Augmented Generation to Summarize and Extract Key Clinical Information From Electronic Health Records," *Journal of Biomedical Informatics* 156 (2024): 104662.

137. E. J. Pérez-Guerrero, I. Mehrotra, S. S. Jain, and M. V. Perez, "Large Language Models as Partners in Medical Literature," *Heart Rhythm* 1 (2024).

138. P. Cheng, X. Yang, A. Chen, et al., "A Study of Generative Large Language Model for Medical Research and Healthcare," *NPJ Digital Medicine* 6, no. 1 (2023): 210.

139. D. Scherbakov, N. Hubig, V. Jansari, A. Bakumenko, and L. A. Lenert, "The Emergence of Large Language Models (Llm) as a Tool in Literature Reviews: An Llm Automated Systematic Review," 2024. arXiv Preprint arXiv:2409.04600.

140. E. Guo, M. Gupta, J. Deng, Y.-J. Park, M. Paget, and C. Naugler, "Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study," *Journal of Medical Internet Research* 26 (2024): e48996.

141. R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large Language Models in Health Care: Development, Applications, and Challenges," *Health Care Science* 2, no. 4 (2023): 255–263.

142. X. Jiang, L. Yan, R. Vavekanand, and M. Hu, "Large Language Models in Healthcare Current Development and Future Directions," *Preprint* (2024), https://doi.org/10.20944/preprints202407.0923.v1.

143. M. Javaid, A. Haleem, and R. P. Singh, "Chatgpt for Healthcare Services: An Emerging Stage for an Innovative Perspective," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 3, no. 1 (2023): 100105.

144. S. Tripathi, R. Sukumaran, and T. S. Cook, "Efficient Healthcare With Large Language Models: Optimizing Clinical Workflow and Enhancing Patient Care," *Journal of the American Medical Informatics Association* 31, no. 6 (2024): 1436–1440.

145. D. Chavali, V. K. Dhiman, and S. C. Katari, "Ai-Powered Virtual Health Assistants: Transforming Patient Engagement Through Virtual Nursing," *Int. J. Of Pharm. Sci* 2 (2024): 613–624.

146. N. Malamas, K. Papangelou, and A. L. Symeonidis, "Upon Improving the Performance of Localized Healthcare Virtual Assistants," *Healthcare (Basel)* 10 (2022): 99.

147. M. Clark and S. Bailey, "Chatbots in Health Care: Connecting Patients to Information," *Canadian Journal of Health Technologies* 4, no. 1 (2024): 1.

148. D. P. Panagoulias, M. Virvou, and G. A. Tsihrintzis, "Augmenting Large Language Models With Rules for Enhanced Domain-Specific Interactions: The Case of Medical Diagnosis," *Electronics (Basel)* 13, no. 2 (2024): 320.

149. H. Fraser, D. Crossland, I. Bacher, et al., "Comparison of Diagnostic and Triage Accuracy of Ada Health and Webmd Symptom Checkers, Chatgpt, and Physicians for Patients in an Emergency Department: Clinical Data Analysis Study," *JMIR mHealth and uHealth* 11, no. 1 (2023): e49995.

150. D. Azevedo, A. Legay, and S. Kieffer, "User Reception of Babylon Health's Chatbot," in *VISIGRAPP (2: HUCAPP)* (SciTePress, 2022), 134–141.

151. Z. Ren, Y. Zhan, B. Yu, L. Ding, and D. Tao, "Healthcare Copilot: Eliciting the Power of General Llms for Medical Consultation," 2024. arXiv Preprint arXiv:2402.13408.

152. D. Lee, "Ai-Based Healthcare Chatbot," *International Journal of Research in Engineering and Technology* 10 (2023): 563–568.

153. A. Ćirković, "Evaluation of Four Artificial Intelligence–Assisted Self-Diagnosis Apps on Three Diagnoses: Two-Year Follow-Up Study," *Journal of Medical Internet Research* 22, no. 12 (2020): e18097.

154. E. Waisberg, J. Ong, M. Masalkhi, et al., "Gpt-4: A New Era of Artificial Intelligence in Medicine," *Irish Journal of Medical Science* 192, no. 6 (2023): 3197–3200.

155. J. Siebielec, M. Ordak, A. Oskroba, A. Dworakowska, and M. Bujalska-Zadrozny, "Assessment Study of Chatgpt-3.5's Performance on the Final Polish Medical Examination: Accuracy in Answering 980 Questions," *Health* 12 (2024): 1637.

156. R. Islam and I. Ahmed, "Gemini-The Most Powerful Llm: Myth or Truth," in *2024 5th Information Communication Technologies Conference (ICTC)* (IEEE, 2024), 303–308.

157. G. K. Gupta, A. Singh, S. V. Manikandan, and A. Ehtesham, "Digital Diagnostics: The Potential of Large Language Models in Recognizing Symptoms of Common Illnesses," 2024. arXiv Preprint arXiv:2405.06712.

158. X. Chen, S. Luo, C.-M. Pun, and S. Wang, "Medprompt: Cross-Modal Prompting for Multi-Task Medical Image Translation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* (Springer, 2024), 61–75.

159. A. S. George, A. S. H. George, A. Shahul, and T. Baskar, "Ai-Driven Breakthroughs in Healthcare: Google Health's Advances and the Future of Medical Ai," *Partners Universal International Innovation Journal* 1, no. 3 (2023): 256–267.

160. S. García-Méndez and F. de Arriba-Pérez, "Large Language Models and Healthcare Alliance: Potential and Challenges of Two Representative Use Cases," *Annals of Biomedical Engineering* 52 (2024): 1–4.

161. I.-C. Hsu and J.-D. Yu, "A Medical Chatbot Using Machine Learning and Natural Language Understanding," *Multimedia Tools and Applications* 81, no. 17 (2022): 23777–23799.

162. A. Babu and S. B. Boddu, "Bert-Based Medical Chatbot: Enhancing Healthcare Communication Through Natural Language Understanding," *Exploratory Research in Clinical and Social Pharmacy* 13 (2024): 100419.

163. D. Inupakutika, D. Akopian, G. Reddy, P. Chalela, S. Kaghyan, and R. Mundlamuri, "Custom Natural Language Understanding for Healthcare Chatbots and a Case Study," in *2024 IEEE International Conference on Digital Health (ICDH)* (IEEE, 2024), 114–122.

164. A. Ait-Mlouk and L. Jiang, "Kbot: A Knowledge Graph Based Chatbot for Natural Language Understanding Over Linked Data," *IEEE Access* 8 (2020): 149220–149230.

165. W. Y. Lin, "Prototyping a Chatbot for Site Managers Using Building Information Modeling (Bim) and Natural Language Understanding (Nlu) Techniques," *Sensors* 23, no. 6 (2023): 2942.

166. S. Bhat Brahmavar, A. Srinivasan, T. Dash, et al., "Generating Novel Leads for Drug Discovery Using Llms With Logical Feedback," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 1 (2024): 21–29, https://doi.org/10.1609/aaai.v38i1.27751, https://ojs.aaai.org/index.php/AAAI/article/view/27751.

167. A. M. Bran, S. Cox, O. Schilter, et al., "Chemcrow: Augmenting Large-Language Models With Chemistry Tools," 2023, https://arxiv.org/abs/2304.05376.

168. S. P. H. Alexander, D. Fabbro, E. Kelly, et al., "The Concise Guide to Pharmacology 2023/24: Enzymes," *British Journal of Pharmacology* 180 (2023): S289–S373, https://api.semanticscholar.org/CorpusID:266376374.

169. H. Hassani, C. Beneki, S. Unger, M. Taj Mazinani, and M. R. Yeganegi, "Text Mining in Big Data Analytics," *Big Data and Cognitive Computing* 4, no. 1 (2020): 1–34.

170. V. Cl, "Big Data and Pharmacovigilance: Data Mining for Adverse Drug Events and Interactions," *P & T* 43, no. 6 (2018): 340–351, https://api.semanticscholar.org/CorpusID:48355575.

171. N. Perera, M. Dehmer, and F. Emmert-Streib, "Named Entity Recognition and Relation Detection for Biomedical Information Extraction," *Frontiers in Cell and Developmental Biology* 8 (2020): 673.

172. Y. Su, X. Jiang, and Y. Chen, "Text Mining Driven Drug-Drug Interaction Detection," in *2013 IEEE International Conference on Bioinformatics and Biomedicine* (IEEE, 2013), 349–354, https://doi.org/10.1109/BIBM.2013.6732517.

173. B. Percha, Y. Garten, and R. B. Altman, "Discovery and Explanation of Drug-Drug Interactions via Text Mining," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 1 (2011): 410–421, https://api.semanticscholar.org/CorpusID:10897158.

174. R. Dhudum, A. Ganeshpurkar, and A. Pawar, "Revolutionizing Drug Discovery: A Comprehensive Review of Ai Applications," *Drugs and Drug Candidates* 3, no. 1 (2024): 148–171, https://doi.org/10.3390/ddc3010009, https://www.mdpi.com/2813-2998/3/1/9.

175. E. Cacace, G. Kritikos, and A. Typas, "Chemical Genetics in Drug Discovery," in *Current Opinion in Systems Biology*, vol. 4 (Elsevier, 2017), 35–42, https://doi.org/10.1016/j.coisb.2017.05.020.

176. T. A. de Oliveira, M. P. da Silva, E. H. B. Maia, et al., "Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods," *Drugs and Drug Candidates* 2, no. 2 (2023): 311–334, https://www.mdpi.com/2813-2998/2/2/17.

177. R. Guha, "On Exploring Structure-Activity Relationships," *Methods in Molecular Biology* 993 (2013): 81–94, https://api.semanticscholar.org/CorpusID:7716005. Humana Press.

178. S. Zhou, Z. Xu, M. Zhang, et al., "Large Language Models for Disease Diagnosis: A Scoping Review," 2024, arXiv Preprint arXiv:2409.00097.

179. M. Omar, D. Brin, B. Glicksberg, and E. Klang, "Utilizing Natural Language Processing and Large Language Models in the Diagnosis and Prediction of Infectious Diseases: A Systematic Review," *American Journal of Infection Control* 52 (2024): 992–1001.

180. Y. Wei, X. Wang, H. Ong, et al., "Enhancing Disease Detection in Radiology Reports Through Fine-Tuning Lightweight Llm on Weak Labels," 2024. arXiv Preprint arXiv:2409.16563.

181. T. Mo, J. Lam, V. Li, and L. Cheung, "Leveraging Large Language Models for Identifying Interpretable Linguistic Markers and Enhancing Alzheimer's Disease Diagnostics," 2024. Medrxiv.

182. Y. Feng, X. Xu, Y. Zhuang, and M. Zhang, "Large Language Models Improve Alzheimer's Disease Diagnosis Using Multi-Modality Data," in *IEEE International Conference on Medical Artificial Intelligence (MedAI)* (IEEE, 2023), 61–66.

183. S. Selva Kumar, A. Khan Mohammed Ajmal Khan, I. A. Banday, M. Gada, and V. V. Shanbhag, "Overcoming Llm Challenges Using Rag-Driven Precision in Coffee Leaf Disease Remediation," in *2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications (ICETCS)* (IEEE, 2024), 1–6.

184. C. Dhavale, T. Pawar, A. Singh, S. Pole, and K. Sabat, "Revolutionizing Farming: Gan-Enhanced Imaging, Cnn Disease Detection, and Llm Farmer Assistant," in *2024 2nd International Conference on Computer, Communication and Control (IC4)* (IEEE, 2024), 1–6.

185. D. Tian, S. Jiang, L. Zhang, X. Lu, and Y. Xu, "The Role of Large Language Models in Medical Image Processing: A Narrative Review," *Quantitative Imaging in Medicine and Surgery* 14, no. 1 (2024): 1108.

186. G. Leonardi, L. Portinale, A. Santomauro, et al., "Enhancing Medical Image Report Generation Through Standard Language Models: Leveraging the Power of Llms in Healthcare," in *2nd AIxIA Workshop on Artificial Intelligence for Healthcare* (IEEE, 2023), 41–45.

187. A. Carriero, L. Groenhoff, E. Vologina, P. Basile, and M. Albera, "Deep Learning in Breast Cancer Imaging: State of the Art and Recent Advancements in Early 2024," *Diagnostics* 14, no. 8 (2024): 848.

188. J. Chamberlin, M. R. Kocher, J. Waltz, et al., "Automated Detection of Lung Nodules and Coronary Artery Calcium Using Artificial Intelligence on Low-Dose Ct Scans for Lung Cancer Screening: Accuracy and Prognostic Value," *BMC Medicine* 19 (2021): 1–14.

189. L. Mosconi, M. Brys, L. Glodzik-Sobanska, S. De Santi, H. Rusinek, and M. J. De Leon, "Early Detection of Alzheimer's Disease Using Neuroimaging," *Experimental Gerontology* 42, no. 1–2 (2007): 129–138.

190. K. Gurucharan Marthi Krishna, A. Chadha, J. Mendola, and A. Shmuel, "Medvisionllama: Leveraging Pre-Trained Large Language Model Layers to Enhance Medical Image Segmentation," 2024. arXiv Preprint arXiv:2410.02458.

191. J. Wu, W. Gan, Z. Chen, S. Wan, and S. Y. Philip, "Multimodal Large Language Models: A Survey," in *2023 IEEE International Conference on Big Data (BigData)* (IEEE, 2023), 2247–2256.

192. Z. Peng, W. Wang, L. Dong, et al., "Grounding Multimodal Large Language Models to the World," in *The Twelfth International Conference on Learning Representations* (2024).

193. D. P. Panagoulias, M. Virvou, and G. A. Tsihrintzis, "Evaluating Llm–Generated Multimodal Diagnosis From Medical Images and Symptom Analysis," 2024. arXiv Preprint arXiv:2402.01730.

194. Y. Li, F. Liang, L. Zhao, et al., "Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-Training Paradigm," 2021. arXiv Preprint arXiv:2110.05208.

195. Z.-Y. Dou, Y. Xu, Z. Gan, et al., "An Empirical Study of Training End-To-End Vision-and-Language Transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022), 18166–18176.

196. C.-C. Lai, C.-Y. Wang, Y.-H. Wang, S.-C. Hsueh, W.-C. Ko, and P.-R. Hsueh, "Global Epidemiology of Coronavirus Disease 2019 (Covid-19): Disease Incidence, Daily Cumulative Index, Mortality, and Their Association With Country Healthcare Resources and Economic Status," *International Journal of Antimicrobial Agents* 55, no. 4 (2020): 105946.

197. S. Stewart, C. Robertson, J. Pan, et al., "Epidemiology of Healthcare-Associated Infection Reported From a Hospital-Wide Incidence Study: Considerations for Infection Prevention and Control Planning," *Journal of Hospital Infection* 114 (2021): 10–22.

198. B. Zhou, G. Yang, Z. Shi, and S. Ma, "Natural Language Processing for Smart Healthcare," *IEEE Reviews in Biomedical Engineering* 1 (2022): 1–12.

199. Z. Al Nazi and W. Peng, "Large Language Models in Healthcare and Medical Domain: A Review," *Inform* 11 (2024): 57.

200. T. Alamo, D. G. Reina, P. Millán Gata, V. M. Preciado, and G. Giordano, "Data-Driven Methods for Present and Future Pandemics: Monitoring, Modelling and Managing," *Annual Reviews in Control* 52 (2021): 448–464.

201. A. Abdulnazar, R. Roller, S. Schulz, and M. Kreuzthaler, "Large Language Models for Clinical Text Cleansing Enhance Medical Concept Normalization," *IEEE Access* 12 (2024): 147981–147990.

202. B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A Survey on Named Entity Recognition—Datasets, Tools, and Methodologies," *Natural Language Processing Journal* 3 (2023): 100017.

203. H. Yin, X. Song, S. Yang, and J. Li, "Sentiment Analysis and Topic Modeling for Covid-19 Vaccine Discussions," *World Wide Web* 25, no. 3 (2022): 1067–1083.

204. S. Lim and R. Schmälzle, "Artificial Intelligence for Health Message Generation: An Empirical Study Using a Large Language Model (Llm) and Prompt Engineering," *Frontiers in Communication* 8 (2023): 1129082.

205. T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu, and G. Haffari, "Continual Learning for Large Language Models: A Survey," 2024. arXiv Preprint arXiv:2402.01364.

206. D. Pissarra, I. Curioso, J. Alveira, et al., "Unlocking the Potential of Large Language Models for Clinical Text Anonymization: A Comparative Study," 2024. arXiv Preprint arXiv:2406.00062.

207. B. Meskó, "The Impact of Multimodal Large Language Models on Health Care's Future," *Journal of Medical Internet Research* 25 (2023): e52865.

208. I. S. Schwartz, K. E. Link, R. Daneshjou, and N. Cortés-Penfield, "Black Box Warning: Large Language Models and the Future of Infectious Diseases Consultation," *Clinical Infectious Diseases* 78, no. 4 (2024): 860–866.

209. F. A. Batarseh, I. Ghassib, D. Chong, and P.-H. Su, "Preventive Healthcare Policies in the Us: Solutions for Disease Management Using Big Data Analytics," *Journal of Big Data* 7, no. 1 (2020): 38.

210. A. Ravi, A. Neinstein, and S. G. Murray, "Large Language Models and Medical Education: Preparing for a Rapid Transformation in How Trainees Will Learn to Be Doctors," *ATS Scholar* 4, no. 3 (2023): 282–292.

211. M. M. Raza, K. P. Venkatesh, and J. C. Kvedar, "Generative Ai and Large Language Models in Health Care: Pathways to Implementation," *NPJ Digital Medicine* 7, no. 1 (2024): 62.

212. W. Betelhem Zewdu, D. Fitsum Gizachew, and W. Michael Melese, "Exploring the Opportunities and Challenges of Implementing Artificial Intelligence in Healthcare: A Systematic Literature Review," *Urologic Oncology: Seminars and Original Investigations* 17 (2023): 4–18.

213. O. A. Farayola, O. L. Olorunfemi, and P. O. Shoetan, “Data Privacy and Security in It: A Review of Techniques and Challenges,” *Computer Science & IT Research Journal* 5, no. 3 (2024): 606–615.

214. Z. Chen, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, “A Survey on Federated Learning,” *Knowledge-Based Systems* 216 (2021): 106775.

215. F. Wu, Z. Li, Y. Li, B. Ding, and J. Gao, “Fedbiot: Llm Local Fine-Tuning in Federated Learning Without Full Model,” in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (ACM, 2024), 3345–3355.

216. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A Survey of Methods for Explaining Black Box Models,” *ACM Computing Surveys* 51, no. 5 (2018): 1–42.

217. Z. Wang, Z. Zhang, A. Traverso, A. Dekker, L. Qian, and P. Sun, “Assessing the Role of Gpt-4 in Thyroid Ultrasound Diagnosis and Treatment Recommendations: Enhancing Interpretability With a Chain of Thought Approach,” *Quantitative Imaging in Medicine and Surgery* 14, no. 2 (2024): 1602–1615.

218. M. Beigi, S. Wang, Y. Shen, et al., “Rethinking the Uncertainty: A Critical Review and Analysis in the Era of Large Language Models,” 2024. arXiv Preprint arXiv:2410.20199.

219. R. AlSaad, A. Abd-Alrazaq, S. Boughorbel, et al., “Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook,” *Journal of Medical Internet Research* 26 (2024): e59505.

220. T. F. Tan, A. J. Thirunavukarasu, J. P. Campbell, et al., “Generative Artificial Intelligence Through Chatgpt and Other Large Language Models in Ophthalmology: Clinical Applications and Challenges,” *Ophthalmology Science* 3, no. 4 (2023): 100394.

221. B. Meskó and E. J. Topol, “The Imperative for Regulatory Oversight of Large Language Models (Or Generative Ai) in Healthcare,” *NPJ Digital Medicine* 6, no. 1 (2023): 120.

222. T. Templin, M. W. Perez, S. Sylvia, J. Leek, and N. Sinnott-Armstrong, “Addressing 6 Challenges in Generative Ai for Digital Health: A Scoping Review,” *PLOS Digital Health* 3, no. 5 (2024): e0000503.

223. L. Banh and G. Strobel, “Generative Artificial Intelligence,” *Electronic Markets* 33, no. 1 (2023): 63.

224. E. Ullah, A. Parwani, M. M. Baig, and R. Singh, “Challenges and Barriers of Using Large Language Models (Llm) Such as Chatgpt for Diagnostic Medicine With a Focus on Digital Pathology–a Recent Scoping Review,” *Diagnostic Pathology* 19, no. 1 (2024): 43.

225. S. Reddy, “Generative Ai in Healthcare: An Implementation Science Informed Translational Path on Application, Integration and Governance,” *Implementation Science* 19, no. 1 (2024): 27.

226. Y. Xia, J. Kim, Y. Chen, et al., “Understanding the Performance and Estimating the Cost of Llm Fine-Tuning,” 2024. arXiv Preprint arXiv:2408.04693.

227. J. T. Reese, D. Danis, J. H. Caufield, et al., “On the Limitations of Large Language Models in Clinical Diagnosis,” 2023. Medrxiv.

228. G. S. Liu, S. Fereydooni, M. C. Lee, et al., “Scoping Review of Deep Learning Research Illuminates Artificial Intelligence Chasm in Otolaryngology-Head and Neck Surgery,” *NPJ Digital Medicine* 8, no. 1 (2025): 1–7.

229. B. Vasey, A. Novak, S. Ather, M. Ibrahim, and P. McCulloch, “Decide-Ai: A New Reporting Guideline and Its Relevance to Artificial Intelligence Studies in Radiology,” *Clinical Radiology* 78, no. 2 (2023): 130–136.

230. L. A. Vaira, J. R. Lechien, V. Abbate, et al., “Validation of the Quality Analysis of Medical Artificial Intelligence (Qamai) Tool: A New Tool to Assess the Quality of Health Information Provided by Ai Platforms,” *European Archives of Oto-Rhino-Laryngology* 281, no. 11 (2024): 6123–6131.

231. L. A. Vaira, J. R. Lechien, V. Abbate, et al., “Enhancing Ai Chatbot Responses in Health Care: The Smart Prompt Structure in Head and Neck Surgery,” *OTO Open* 9, no. 1 (2025): e70075.