BioMedInformatics

Review

Recent Advances in Large Language Models for Healthcare

Khalid Nassiri and Moulay A. Akhloufi *

Perception, Robotics and Intelligent Machines (PRIME), Department of Computer Science, Université de Moncton, Moncton, NB E1A 3E9, Canada; khalid.nassiri@umoncton.ca

* Correspondence: moulay.akhloufi@umoncton.ca

Abstract: Recent advances in the field of large language models (LLMs) underline their high potential for applications in a variety of sectors. Their use in healthcare, in particular, holds out promising prospects for improving medical practices. As we highlight in this paper, LLMs have demonstrated remarkable capabilities in language understanding and generation that could indeed be put to good use in the medical field. We also present the main architectures of these models, such as GPT, Bloom, or LLaMA, composed of billions of parameters. We then examine recent trends in the medical datasets used to train these models. We classify them according to different criteria, such as size, source, or subject (patient records, scientific articles, etc.). We mention that LLMs could help improve patient care, accelerate medical research, and optimize the efficiency of healthcare systems such as assisted diagnosis. We also highlight several technical and ethical issues that need to be resolved before LLMs can be used extensively in the medical field. Consequently, we propose a discussion of the capabilities offered by new generations of linguistic models and their limitations when deployed in a domain such as healthcare.

Keywords: large language models (LLMs); transformers; medical datasets; foundation models

check for updates

Citation: Nassiri, K.; Akhloufi, M.A. Recent Advances in Large Language Models for Healthcare. *BioMedInformatics* **2024**, *4*, 1097–1143. https://doi.org/10.3390/biomedinformatics4020062

Academic Editors: Ognjen Arandjelović and Alexandre G. De Brevern

Received: 25 January 2024
Revised: 24 February 2024
Accepted: 25 March 2024
Published: 16 April 2024

1. Introduction

Recent advances in the field of artificial intelligence (AI) have enabled the development of increasingly powerful linguistic models capable of generating text fluently and coherently. Among these models, "large language models" (LLMs) stand out for their imposing size and their ability to learn enormous amounts of textual data. Models like GPT-3.5 [1] and GPT-4 [2] developed by OpenAI [3], or Bard, created by Google [4], have billions of parameters and have demonstrated impressive comprehension skills and language generation.

These fascinating advancements in natural language processing (NLP) have promising implications in many fields, including healthcare [5–7]. Indeed, they offer new perspectives for improving patient care [8,9], accelerating medical research [10,11], supporting decision-making [12,13], accelerating diagnosis [14,15], and making health systems more efficient [16,17].

These models could also provide valuable assistance to healthcare professionals by helping them interpret complex patient records and develop personalized treatment plans [18] as well as manage the increasing amount of medical literature.

In the clinical domain, these models could help make more accurate diagnoses by analyzing patient medical records [19]. They could also serve as virtual assistants to provide personalized health information or even simulate therapeutic conversations [20–22]. The automatic generation of medical record summaries or examination reports is another promising application [23–25].

For biomedical research, the use of extensive linguistic models paves the way for rapid information extraction from huge databases of scientific publications [26,27]. They can also generate new research hypotheses by making new connections in the literature. These applications would significantly accelerate the discovery process in biomedicine [28,29].

Additionally, these advanced language models could improve the administrative efficiency of health systems [30]. They would be able to extract key information from massive medical databases, automate the production of certain documents, or even help in decision-making for the optimal allocation of resources [31–33].

LLMs such as GPT-3.5, GPT-4, Bard, LLaMA, and Bloom have shown impressive results in various activities related to clinical language comprehension. Nevertheless, it is essential to evaluate them in detail in the medical context, which is characterized by its distinct nuances and complexities compared to common texts.

It is therefore essential to continue studies in order to verify the capacity of these innovative linguistic models in real clinical situations, whether for decision-making, diagnostic assistance, or the personalization of treatments. Rigorous evaluation is the key to taking full advantage of this technology and transforming medicine through better mastery and use of clinical language.

Although promising, the use of this AI in health also raises ethical and technological challenges that must be addressed. However, their potential for accelerating medical progress and improving the quality of care seems immense. The coming years will tell us to what extent these revolutionary models are capable of transforming medicine and health research.

The main contributions of this paper are the following:

- We analyze major large language model (LLM) architectures such as ChatGPT, Bloom, and LLaMA, which are composed of billions of parameters and have demonstrated impressive capabilities in natural language understanding and generation.
- We present recent trends in the medical datasets used to train such models. We classify these datasets according to different criteria, such as their size, source (e.g., patient files, scientific articles), and subject matter.
- We highlight the potential of LLMs to improve patient care through applications like assisted diagnosis, accelerate medical research by analyzing literature at scale, and optimize the efficiency of health systems through automation.
- We discuss key challenges for practically applying LLMs in medicine, particularly important ethical issues around privacy, confidentiality, and the risk of algorithmic biases negatively impacting patient outcomes or exacerbating health inequities. Addressing these challenges will be critical to ensuring that LLMs can safely and equitably benefit public health.

As depicted in Figure 1, this study is deployed according to a well-defined architecture that aims to enlighten the reader on the different facets of LLMs and their relevance in medical diagnoses. We will begin, in Section 2, our exploration with a brief history of LLMs. Section 3 serves as an introduction to the transformer architecture, laying the foundation for understanding the following sections. Building on this foundation, in Section 4, we will delve deeper into the specific architecture of LLMs. Section 5 illustrates the practical applications of LLMs, with an emphasis on their use in medical diagnosis. Section 6 presents a comprehensive review of medical datasets, segmenting them into three key categories. Section 7 focuses on the major innovation, which is the advent of foundation models in the AI landscape, highlighting their relevance in the clinical domain. Before concluding this article in the final Section 9, Section 8 offers critical reflections, assessing both the benefits of LLMs and not neglecting the challenges inherent in them.

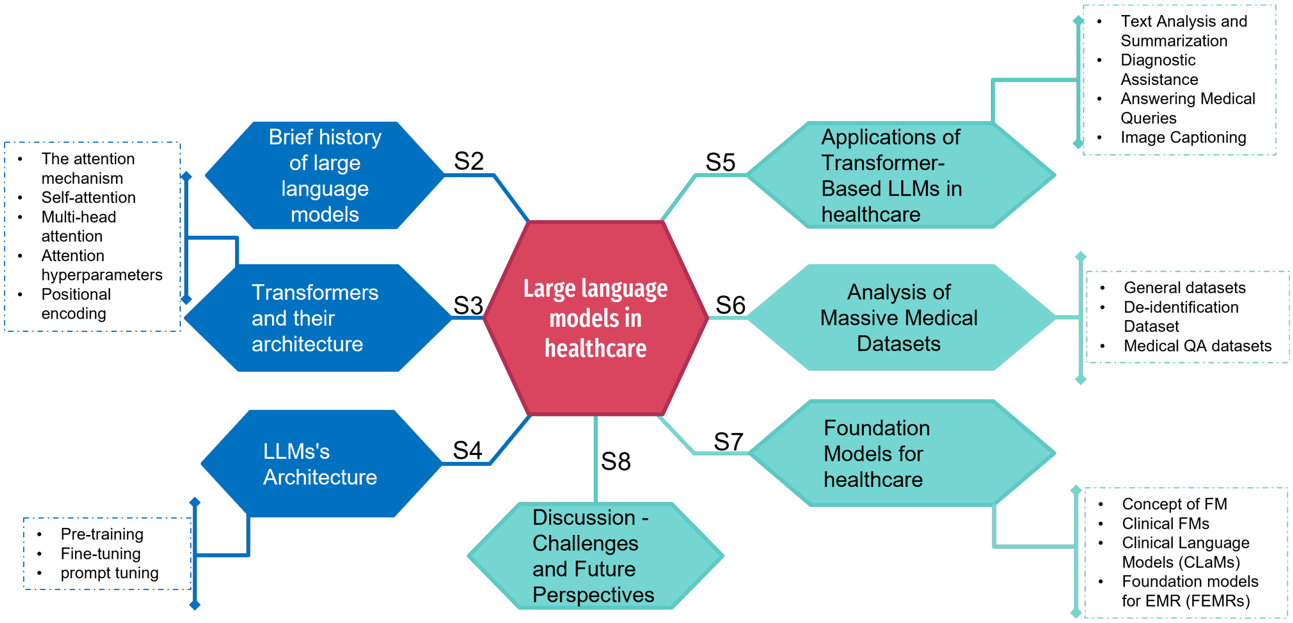
Large language models in healthcare
S2
Brief history of large language models
S3
Transformers and their architecture
- The attention mechanism
- Self-attention
- Multi-head attention
- Attention hyperparameters
- Positional encoding
S4
LLMs's Architecture
- Pre-training
- Fine-tuning
- prompt tuning
S5
Applications of Transformer-Based LLMs in healthcare
- Text Analysis and Summarization
- Diagnostic Assistance
- Answering Medical Queries
- Image Captioning
S6
Analysis of Massive Medical Datasets
- General datasets
- De-identification Dataset
- Medical QA datasets
S7
Foundation Models for healthcare
- Concept of FM
- Clinical FMs
- Clinical Language Models (CLaMs)
- Foundation models for EMR (FEMRs)
S8
Discussion - Challenges and Future Perspectives

Figure 1. Structural diagram presenting the topics covered in our paper.

2. Brief History of Large Language Models

The first language models were n-gram models [34], which estimate the probability of a word based on previous n − 1 words. They began to be used in the 1980s and are still used today. However, they do not capture the semantics of the language well.

During the 2000s, researchers introduced topic models like latent Dirichlet allocation (LDA) [35]. These models have the ability to detect themes within large collections of text and are particularly useful for analyzing large amounts of health-related textual data.

Language models based on neural networks, such as word2vec [36] and GloVe [37], began to emerge in the 2010s. They learn vector representations of words that capture semantic and syntactic relationships. They have been used for tasks such as extracting information from clinical texts.

The introduction of the transformer architecture in 2017 brought a revolution in the field of NLP, paving the way for the emergence of large-scale pre-trained models such as BERT [4], ELMo [38], RoBERTa [39], and GPT-3 [3]. These models are trained on massive amounts of general domain text and can then be fine-tuned for specific healthcare applications. They have been used for tasks such as named entity recognition, sentiment analysis, and question answering (QA) in clinical text. Indeed, domain-specific versions of BERT such as BioBERT [40] and ClinicalBERT [41] have been developed to address clinical language comprehension tasks.

More recently, LLMs have continued to evolve, demonstrating cutting-edge performance in all fields, including healthcare. They are being applied in new ways in healthcare, such as to facilitate clinical documentation, identify adverse drug reactions, and predict health outcomes from patient notes. However, the specialized clinical vocabularies, acronyms, and abbreviations present in the text remain a challenge.

Over time, LLMs have steadily increased in size and performance, such as GPT-3.5 and GPT-4, as well as Bard. This opened the way to new use cases: more comprehensive virtual assistants [42], integration with patient files [43], diagnostic/therapeutic recommendations [44], etc.

Today, research is exploring many avenues: personalized medicine with Omics data [45], medical image analysis [46], clinical decision support [47,48], comprehensive healthcare assistants [49,50], and biomedical knowledge bases [51,52], etc.

While LLMs have tremendous potential, their development raises major ethical challenges: guaranteeing patient safety, combating bias, verifying and explaining recommendations, respecting privacy, and ensuring complementarity with caregivers. Researchers are working actively on these issues so that AI benefits the healthcare system responsibly.

3. Transformers and Their Architecture

Transformers, a category of deep learning (DL) models, are predominantly employed for tasks related to NLP. These models were first introduced by Google in 2017 through a paper titled "Attention is All You Need" [53]. The fundamental element of transformers is the attention mechanism (see Section 3.1). This mechanism enables the model to comprehend the contextual associations between words (or other elements) within a sentence.

Transformers use multi-head self-attention to analyze the relationship between words in a sentence. Self-attention means that words attend to their relationship with other words in the same sequence, without regard to their relative or absolute position.

The basic architecture of transformers consists of an encoder and a decoder. The encoder helps process the input sequence, while the decoder generates the output sequence. Common transformer architectures include BERT [4], GPT [54], T5 [55], etc. BERT (Bidirectional Encoder Representations from Transformers) introduced bidirectional training, which looks at the context from both left and right. GPT (Generative Pre-trained Transformer) models like GPT-2 [56], GPT-3 [3], and GPT-4 [2] are able to generate new text. T5 (Text-To-Text Transfer Transformer) can perform a wide range of text-based tasks like summarization, QA, and translation [57].

All transformer models follow the same basic structure—embedding, encoding, and decoding. However, they differ in pre-training objectives, model size, number of encoder/decoder stacks, attention types, etc. Later models like T5, GPT-3, and GPT-4 have billions of parameters to handle more complex tasks through transfer learning from massive text corpora.

3.1. The Attention Mechanism

The attention mechanism aims to palliate the loss of information transmitted to the decoder, as it is only the hidden state created during the last phase by the encoder that is provided as input to the decoder.

The original work of Larochelle and Hinton [58] introduced this approach to the field of computer vision. By analyzing several regions of an image separately, i.e., by considering different extracts, it is possible for a learning algorithm to gradually accumulate knowledge about the shapes and objects present. By analyzing each segment in turn, the model can build up a global understanding of the image as a whole, which will ultimately enable it to assign a relevant category to it. Authors initially proposed this method, which involves examining the various parts of an image and, after assimilating the details of each, arriving at a precise classification.

According to Equation (1), the attention mechanism, recommended by [59,60], has played a crucial role in improving the performance of machine translation systems. This approach offers the model the ability to focus on essential segments of the input sequence.

The main idea behind attention is to evaluate the relationship between parts of two different sequences. In NLP, especially in a sequence-to-sequence framework, the attention mechanism aims to signal to the model which word in sequence "B" should be privileged in relation to a specific word in sequence "A".

A model with attention differs from a classic seq2seq model in two major respects: Firstly, instead of only transmitting the ultimate hidden state from the encoder to the decoder, it transmits all the hidden states to the decoder, thus enriching the information transmitted. Secondly, before producing its output, the decoder integrates an additional step. It evaluates all the hidden states received from the encoder, assigning them a score via multiplication by their softmax value, to better target the crucial elements of the input.

For each word in the input sequence, a contextualized attention layer (self-attention) generates a vector representative of its importance (attention vector). Although explanations are presented here as vectors for simplicity's sake, calculations actually involve matrix representations (several vectors juxtaposed), with each word associated with a row of a matrix. Thus, attention vectors actually describe the relative influence of each row of a matrix representing the sequence as a whole, allowing contextual interactions to be captured at all levels.

To calculate the attention vector for each encoded input, we need to consider three types of vectors:

- The query vector, q, represents the encoded sequence up to that point.
- The key vector, noted k, corresponds to a projection of the encoded entry under consideration.
- The value vector, v, contains the information relative to this same encoded entry.

Thus, to obtain the attention vector associated with a given encoder input, the mechanism takes into account these three vectors: q, k, and v, respectively, the request vector, key vector, and value vector.

These vectors q, k, and v are indexed by the position of the word in the processed sequence. For example, for the first word, we will have the vectors q_1, k_1, and v_1. They are generated by projecting the x vector representation of each word using three matrices:

- The Q matrix produces the q query vectors.
- Matrix K produces key vectors k.
- Matrix V generates value vectors v.

These three matrices, Q, K, and V are learned during the training process of the transformer model [53] so as to optimize the calculation of contextual attention. The vectors q, k, and v for each word are thus derived by multiplying the vector representation x with these matrices.

$$Attention(Q, V, K) = softmax(\frac{QV^T}{\sqrt{d_k}})V \tag{1}$$

d_k is the hidden dimensionality for keys. V^T is the transpose of a V matrix.

3.2. Self-Attention

Self-attention aims to model the contextual relationships existing within a single sequence thanks to the attention mechanism [53]. Within a self-attention layer, we seek to determine the interdependence between different elements (words), in order to obtain a vector representation enriched by the global context. Unlike traditional attention between input and output sequences, here, we calculate attention scores between the elements of the single sequence under consideration. This internal self-attention is an essential building block in the transformer architecture, enabling words to be enriched by their linguistic environment within the same sentence. Self-attention is also called intra-attention [61–63].

3.3. Multi-Head Attention

The multi-head attention architecture is a major advantage of the transformer model. By dividing calculations between several "heads" carrying out their processing in parallel, it considerably speeds up processing, since each head handles part of the data simultaneously. Parallelization also gives the system greater modeling capacity.

Each head can then model contextual relationships from its own angle or scale, enriching the final representation with complementary perspectives. Multi-headed attention thus captures dependencies in a finer, more complex way between the various elements processed. Furthermore, integrating a variety of contexts enables each word to be better represented.

This architecture significantly increases the flexibility and expressive capacity of the transformer model without critically increasing the number of parameters. It also makes the system more robust, as each head can compensate for any local errors of the others. All in all, multi-headed attention makes the most of the possibilities offered by parallel computing to enhance contextual modeling.

3.4. Attention Hyperparameters

The dimensions of the data manipulated by the transformer model are configured using three hyperparameters:

- The embedding size, which corresponds to the dimension of the vectors used to represent the input elements (words and tokens). Present throughout the model, this dimension also defines its capacity, called "model size".
- The size of the queries (equal to that of the keys and values), i.e., the dimension of the vectors produced by the three linear layers generating the matrices of queries, keys, and values required for attentional calculations.
- The number of attentional heads, which determines the number of attentional processing blocks operating in parallel.

These three hyperparameters determine the vector formats used at each stage of the transformer, from input to output. They directly condition its expressive capabilities and the volume of data it can process in a sophisticated way.

3.5. Positional Encoding

As the transformer model has neither recurrence nor convolution, it has no intrinsic mechanisms for exploiting the order of elements within input sequences. This property is essential for many tasks, such as translation or text comprehension. To remedy this, positional embeddings are added to the initial token plunges.

In concrete terms, each position in the sequence is assigned a vector encoding its relative or absolute place. These positional embeddings, which have the same dimensions as the usual embeddings, can then simply be added to them. In this way, the model has additional information on the position of each token within the sequence.

These positional embeddings can be learned during training as free parameters, or fixed using mathematical functions such as sine/cosine. Placed at the input of the encoder and decoder layers, they enable the latter to exploit the missing sequential dimension, significantly improving performance on many tasks. Their use has proved particularly beneficial in initial transformer models whose design rejects recurrence and convolution [64].

3.6. Transformers

The development of the digital world would not have been possible without advancements in automatic NLP. However, for a long time, NLP techniques remained limited due to insufficient progress in AI.

Recurrent neural networks (RNN) [61,63,65], and convolutional neural networks (CNN) [53], widely used in the past for NLP tasks, had the disadvantage of processing text sequentially. This approach proved inefficient with massively parallel computing units such as GPUs, which became essential with the advent of DL on large datasets.

It was against this backdrop that the transformer model emerged in 2017 by Vaswani et al. [53] of Google Brain and Google Research. Based on attention as a basic primitive, it enabled sequences to be processed in a truly parallel way for the first time, thanks to multi-headed calculations.

In terms of performance, transformer proved superior to RNN and CNN on natural language understanding and generation tasks. It was able to learn more efficiently and in parallel, better capturing long-range dependencies. Its excellent evaluation results have made the transformer an essential component of modern NLP.

It has revolutionized the field by enabling language processing on a very large scale, paving the way for major advancements such as human-quality machine translation.

Encoders and decoders are the two essential components of the original transformer model (see Figure 2). Each part is composed of a stack of N = 6 layers that are all identical. The output of one layer is the input of the next layer until the final representation (prediction) is reached [53].

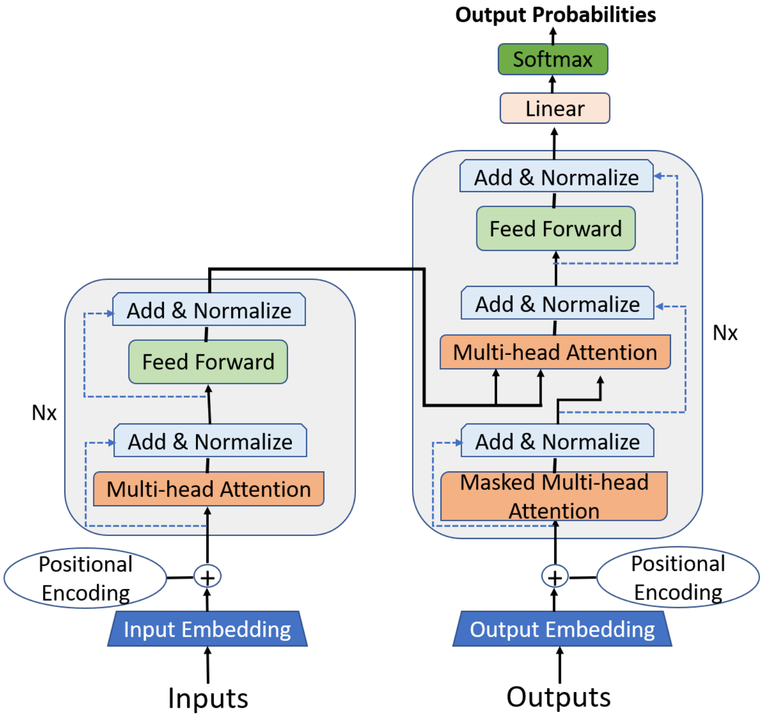

Figure 2. The transformer model architecture.

Transformer models, as represented in Figure 2, are distinguished by the presence of two fundamental elements: encoders and decoders. These two components form the basis of the architecture. Each of these elements is designed around a set of six identical layers, where the output of each layer feeds the input of the next, enabling a progressive transformation of information until the final prediction is formulated.

The encoder, which forms the left-hand side of this architecture, is structured into two main blocks, both of which are neural networks.

A self-attention layer works to preserve the dependencies between words in a sequence. It analyzes each word in relation to the others, in order to identify its context. A feed-forward neural network, which performs complex transformations on the data received from the self-attention layer. Its main task is to transform an input sequence into a series of continuous representations, which then serve as inputs for the decoder.

The decoder, located on the right-hand side of the architecture, is structured in a similar way as the encoder but also incorporates an "Encoder-Decoder Attention" layer. The latter plays a crucial role: it facilitates the attention mechanism between the input sequence (once encoded) and the output sequence (during its decoding phase). The aim is to ensure that each word in the output sequence takes account of all the words in the input sequence.

The decoding process is based on the combination of the encoder output and the output generated by the decoder during the previous time step, enabling the output sequence to be built up progressively.

One of the main strengths of transformer architectures such as LLMs lies in their ability to extract essential intermediate features, such as syntactic structures and semantic integration of words. This ability enables LLMs to represent human language knowledge accurately and richly, making them powerful in a variety of tasks such as QA, sentiment classification, and machine translation. Moreover, LLMs are versatile, as they can be adapted to new challenges by reusing the same pre-trained model, which often puts them head and shoulders above previous methods. Their ability to extract and exploit relevant linguistic information makes LLMs powerful and versatile tools in NLP.

In the following section, we explain the architecture of LLMs in detail, highlighting the key mechanisms that contribute to their interesting results in text understanding and generation.

4. LLMs' Architecture

LLMs have made significant advancements in recent years. First, the transformer models introduced a novel attention-based neural architecture for NLP.

These advanced models, with their billions of parameters, have become feasible thanks to recent advancements in computational capabilities and model architecture [66], as shown in Figure 3. For example, GPT-3, which relies on extensive data, has around 175 billion parameters [67]. Meanwhile, the open-source LLaMA model series ranges between 7 and 70 billion parameters [68,69]. Table 1 summarizes the main LLMs, giving their general architecture as well as an indication of the number of parameters in some of the basic versions used in research. This enables a quick comparison of the size and specific features of each model.

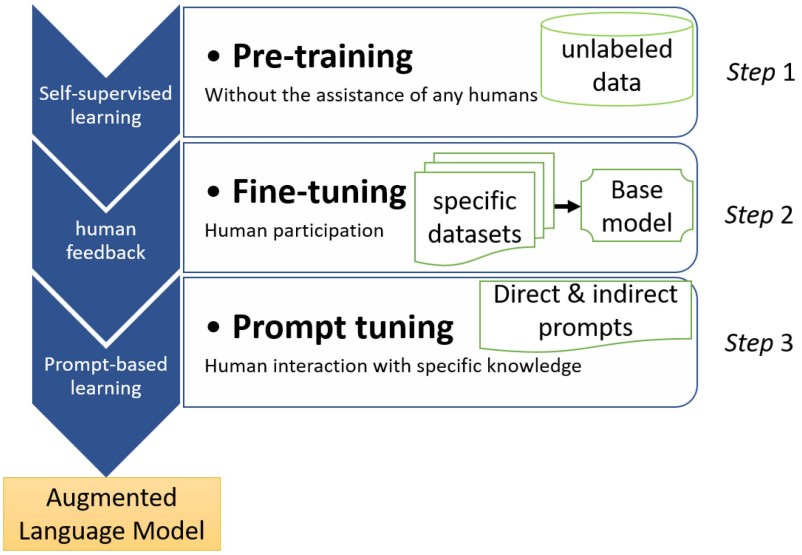

Figure 3. The LLM architecture.

Table 1. LLM architectures and their characteristics.

Architectures	Models	Parameters
Autoencoders (Encoder Only)	BERT RoBERTa	110 M 125 M
Autoregressors (Decoder Only)	GPT-3 GPT-4 Bloom PaLM LLaMA StableLM	175 B 1.76 T 176 B 540 B from 7 to 70 B from 15 to 65 B
Sequence-to-Sequence (Encoder–Decoder)	T5 BART	220 M 139 M

The initial phase of LLM training, called **pre-training**, uses a self-supervised method using large amounts of unlabeled data, which include sources such as general web content, Wikipedia, Github repositories, social media, and BooksCorpus [68,70]. The main goal of training is to predict the next word in a series, which requires significant resources [71]. This involves converting the text into tokens before entering it into the template [72]. The outcome of this process is a base model designed primarily for basic language generation but unable to perform complex tasks.

After initial pre-training, the model undergoes a **fine-tuning** phase to specialize it for certain tasks [72]. Here, the model can be trained on narrower domain-specific datasets, like medical records for healthcare applications.

This fine-tuning can be improved through various techniques. A constitutional AI approach embeds predefined rules or principles directly into the model architecture [73]. Reward-based training involves human evaluators assessing the quality of multiple model outputs to provide feedback [70,74]. Reinforcement learning based on human feedback uses a comparison-based system to optimize responses through iterative human feedback [74].

This fine-tuning step requires less computational power but more human resources compared to pre-training. It customizes the model to perform a specific task, like a chatbot, with controlled, targeted results. The fine-tuned model resulting from this phase is then deployed for flexible applications in its domain of specialization.

The goal is to leverage general pre-training and then fine-tune the model's capabilities for the nuances of its intended use case through various focused training approaches during fine-tuning. The result is a model that is both adapted and versatile for its specialist field of application.

LLMs have an impressive ability to adapt to unfamiliar tasks and demonstrate remarkable reasoning skills [75,76]. However, to fully exploit their potential in specialized fields such as medicine, more specific training strategies are essential. These strategies could encompass direct prompting methods such as few-shot learning [3], in which a limited set of task examples during testing guides the model's results, and zero-shot learning [77], in which the model operates without any prior specific examples. In addition, refined techniques such as chain-of-thought prompting [78], which prompts the model to sequentially decompose its reasoning, and self-consistency checks [79], which ask the model to confirm the consistency of its responses, are also crucial.

Reinforcement learning (RL) is also a complementary technique in the fine-tuning process of LLMs, aimed at improving and better aligning pre-trained models. To enhance the performance of these LLMs, optimization methods inspired by RL or directly derived from RL are implemented. Notable examples include RL from human feedback (RLHF), direct preference optimization (DPO) and proximal policy optimization (PPO).

RLHF [2,69] integrates RF techniques with human feedback to refine the language model. DPO, introduced by Rafailov et al. [80], focuses on direct preference optimization for model-generated responses. PPO, initially conceived by Schulman et al. [81] and later adapted by Tunstall et al. [82], employs proximal policy optimization for LLM fine-tuning.

These RL-based approaches have proven effective, particularly when combined with instruction tuning, in improving the relevance and quality of responses generated by LLMs. However, it is important to note that, like instruction tuning, these methods are primarily aimed at improving the quality of responses and their behavioral appropriateness, without necessarily increasing the breadth of knowledge that the model can demonstrate.

Instruction **prompt tuning**, developed by Lester et al. [83], is a promising technique for efficiently updating model parameters, enhancing its performance on many downstream medical tasks. This approach has advantages over prompt methods in a few examples. In general, these methods enrich the initial model fine-tuning processes, enhancing their suitability for medical problems. A recent example is the Flan-PaLM model [84]. Figure 4 compares classical fine-tuning and prompt tuning for adapting pre-trained language models to specific tasks. In classical fine-tuning, each task requires the creation of a dedicated model, leading to multiple versions of the model with different configurations for each task. This approach is complex and requires significant resources. Prompt tuning, on the other hand, simplifies the adaptation process. Instead of creating separate models, we simply provide the original pre-trained model with a "prompt" specific to each task. This "prompt" acts as an instruction that allows the model to handle different tasks without requiring major modifications.

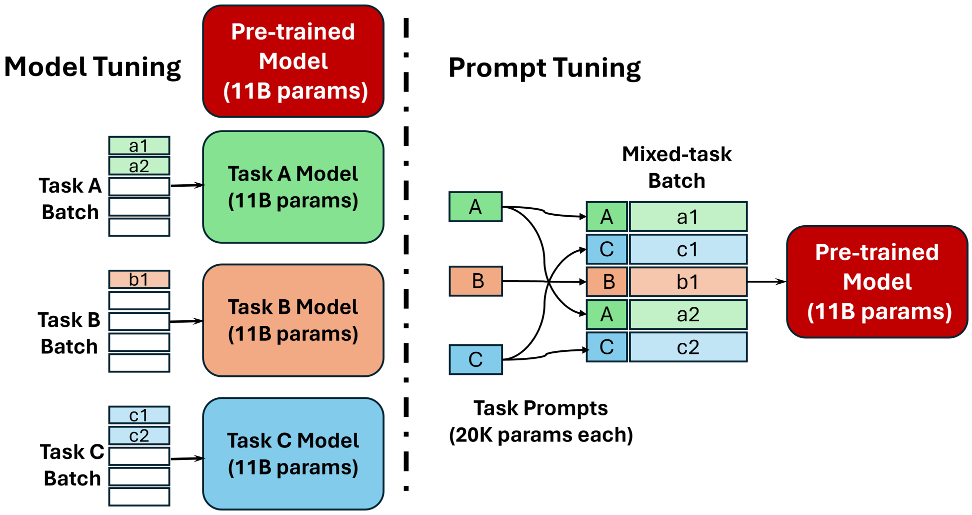
Model Tuning
Pre-trained Model (11B params)
Task A Batch
a1
a2
Task A Model (11B params)
Task B Batch
b1
Task B Model (11B params)
Task C Batch
c1
c2
Task C Model (11B params)
Prompt Tuning
Mixed-task Batch
A
B
C
A a1
C c1
B b1
A a2
C c2
Pre-trained Model (11B params)
Task Prompts (20K params each)

Figure 4. Differences between classic fine-tuning and prompt tuning.

The major advantage of prompt tuning lies in its efficiency. For example, with a large T5 model [55], traditional fine-tuning for each task can require around 11 billion parameters, which is equivalent to creating new models for each task. In contrast, prompt tuning only requires about 20,480 parameters per task "prompt", representing a significant reduction of five orders of magnitude. It is therefore possible to use a single adaptable model with minimal modifications to accomplish various tasks.

Consequently, prompt tuning allows for the adaptation of a single pre-trained model to a multitude of tasks, significantly reducing the computational resources required and improving scalability compared to traditional fine-tuning. This breakthrough paves the way for more widespread use of language models in various domains.

Understanding training methods and their ongoing evolution provides essential insights for assessing the current capabilities of models and identifying their potential future applications. In specialized fields such as healthcare, where precision and expertise are essential, an in-depth analysis of model specialization techniques is crucial to assessing their full potential.

Among these techniques, retrieval augmented generation (RAG) was introduced by Lewis et al. [85]. This method aims to enhance the capabilities of LLMs, especially in knowledge-intensive tasks, by integrating external sources of information. Originally, the application of RAG required task-specific training. However, subsequent research [86] revealed that a pre-trained model incorporating the RAG method could increase its performance without the need for additional training.

The fundamental principle of RAG is based on the use of an auxiliary knowledge base and an input query. The RAG architecture is then used to identify relevant documents in this database that are close to the query. These retrieved documents are then merged with the initial query, providing the model with an enriched, in-depth context on the queried subject.

The success of LLMs is attributed to their ability to generalize from the massive amounts of data that they are exposed to during training. This enables them to perform a wide array of language-related tasks with impressive accuracy, including text completion, language translation, sentiment analysis, and more. LLMs have also demonstrated proficiency in understanding context and generating contextually appropriate responses in conversational settings. Among the instances of LLMs, notable examples include PaLM [87], GPT-3 [3], LLaMA [68], PaLM2 [88] utilized in the BARD chatbot, Bloom [89], and GPT-4 [2]. These models are trained on billions of tokens sourced from datasets such as Common Crawl, WebText2, Wikipedia, Stack Exchange, PUBMED, ArXiv, Github, and many other diverse repositories.

Additionally, researchers are exploring, using LLMs, the creation of systems capable of understanding and applying complex instructions from text. In this context, models such as Alpaca [90], StableLM [91], and Dolly [92] stand out. Alpaca, based on Meta's LLaMA 7B model, showed that a lighter model can compete with heavier models like OpenAI's text-DaVinci-003. In addition, Alpaca is more economical and easy to reproduce thanks to its open structure.

This indicates that it is entirely possible for less demanding models to compete in performance while still being transparent and accessible, thanks to their free nature. StableLM and Dolly are also working in this direction, seeking to better respond to requests formulated in natural language.

LLMs now have trillions of learning data tokens and the number of their parameters has grown rapidly [68]. Rather than improvements in architectural design, the amount of data, parameters, and computational resources appear to be a driving factor in the capabilities of LLMs [93].

While LLMs have opened up new possibilities for understanding and generating natural language, they also pose challenges and ethical considerations. The risk of bias in learning data, ethical concerns relating to content generation, and issues of data privacy are some areas that researchers and practitioners continue to address.

4.1. ChatGPT

ChatGPT, developed by OpenAI and based on the GPT architecture [54], was officially released in June 2020. This model, which has undergone several updates, has been trained using a vast corpus of textual sources such as books, web content, and articles, giving it the ability to capture the complexities of human language.

ChatGPT is able to produce texts that reflect human articulation in response to specific prompts. This prowess makes it invaluable in a variety of NLP applications, from chatbots to translation to text summarization. Furthermore, its adaptability means that it can be fine-tuned for particular tasks using more concentrated datasets.

To explore its origins, the first GPT-1 [54] used a 40 GB text dataset for training. In contrast, its successors—GPT-2 [56], GPT-3 [3], and GPT-4 [2]—benefited from much larger datasets. This progressive increase in training data propelled ChatGPT's effectiveness, with GPT-4's results becoming practically indistinguishable from human-written text. Table 2 illustrates the evolution of the different versions of GPT, specifically in the context of the healthcare sector.

Table 2. Evolution of GPT versions in the healthcare sector.

Version	Year	Number of Parameters	Healthcare Applications	Advances/Particularities in Healthcare
GPT	2018	110 M	Basic medical text generation.	A leader in automatic NLP, but limited in its ability to understand complex medical terminology.
GPT-2	2019	1.5 Bn	Improvements in medical text generation, first drafts of medical article summaries.	Improved consistency in text generation, but still with limits in terms of precision and specific medical context.
GPT-3	2020	175 Bn	Search summaries, assistance in interpreting medical data, generation of medical QA.	Enhanced ability to understand and generate complex medical texts, useful for literature analysis and health consulting applications.
GPT-4	2023	Unknown	In-depth analysis of clinical cases, integration into assisted diagnosis systems, personalized health advice.	Significant advancement in language understanding and accuracy, and the ability to integrate textual and visual data for more comprehensive analyses.

Each GPT version has marked a significant advancement over the previous one, with improvements in terms of model complexity, language understanding, and polyvalence of application domains. GPT-4's capability to manage multimodal inputs (text and images) is a notable evolution over previous versions [2].

Researchers have conducted investigations to assess the usefulness of ChatGPT and InstructGPT [74] for healthcare, as well as their suitability for specific medical tasks. For example, Luo et al. [94] developed BioGPT, a model based on the GPT-2 framework pre-trained on 15 million PUBMED abstracts. BioGPT outperformed other state-of-the-art models in a range of tasks, including QA, relationship extraction, and document classification. Likewise, BioMedLM 2.7B (formerly called PUBMEDGPT [95]), which is pre-trained on both PUBMED abstracts and full text, illustrates ongoing progress in this area. In addition, researchers have used GPT-4 to create multimodal medical LLMs, and early results are promising in this area.

Current research in the biomedical sector focuses mainly on the use of ChatGPT for data assistance. These approaches train models of small size using ChatGPT's distilled or translated knowledge. A notable example is Chatdoctor [32], which marks the first initiative to adapt language and machine learning (LLM) models to the biomedical domain. They fine-tuned the LLaMa model through dialogue simulations generated via ChatGPT.

Another example is DoctorGLM [96], which uses ChatGLM-6B as a base model and fine-tunes it using the Chinese translation of the ChatDoctor dataset, which was obtained using ChatGPT. In addition, Chen et al. [97] developed an improved Chinese and medical linguistic model in their LLM collection.

These works collectively demonstrate the potential of LLMs to be successfully applied in the biomedical domain. They highlight the adaptability of LLMs to meet the specific needs of medical assistance and underline the importance of using synthesized and translated data to form specialized models in this field.

4.2. PaLM

The Pathways Language Model (PaLM), introduced in 2019 [87], is a powerful architecture built upon the transformer decoder, a densely connected language model. PaLM stands out from other models as it prioritizes text generation over input processing, resulting in exceptional efficiency for tasks like generating text. Google developed the Pathways approach specifically for training large-scale machine learning models, and PaLM benefits from this methodology. Pathways enables flexible and scalable model training, a critical factor in handling the immense volumes of data necessary to train such models.

PaLM has been trained on a massive dataset consisting of 780 billion tokens. This dataset includes a variety of textual sources, such as web pages, Wikipedia articles, source code, conversations on social networks, press articles, and books. By incorporating these different sources, PaLM is exposed to various languages, knowledge, and text styles. This enables it to acquire a thorough understanding of language, ranging from structured and specialized knowledge to informal expressions and literary narratives.

With 540 billion parameters, PaLM has a significant ability to model complex relationships, which is advantageous for analyzing and synthesizing detailed medical information. It can potentially be used to generate and summarize medical information, help interpret health data, and answer general medical questions [98]. One of its technical innovations, the "Chain-of-Thought Prompting" method [78], enables it to process complex queries in several stages, which could prove beneficial for solving complex medical problems or interpreting health data in several phases.

A standout feature of PaLM is its competence in addressing out-of-vocabulary (OOV) words, which are terms not seen during its training phase. Rather than stumbling over these unfamiliar words, PaLM can generate fitting contextual replacements, thereby elevating its overall performance in text generation.

Researchers first adapted the PaLM model to medical QA, leading to the creation of [84]. This model established state-of-the-art benchmarks for QA. Based on these results, the Med-PaLM model was introduced using instruction tuning, proving its effectiveness in areas such as clinical expertise, scientific consensus, and medical reasoning processes [99].

This approach has been extended to develop a multimodal medical LLM. These PaLM-centric models highlight the benefits of adapting basic models to the specific needs of medical scenarios. Flan-PaLM is a specially designed version of PaLM for processing instructions [84]. The latest version, called Med-PaLM2, was unveiled at Google Health's annual event [98]. This version was developed based on PaLM2, which is the fundamental language model used by Google's chatbot, Bard.

Recently, Google developed a multimodal model called Med-PaLM M [100], which follows on from its PaLM-E vision-language model [101], and has the ability to synthesize and communicate information from medical images such as chest X-rays, dermatology images, pathology slides, and other biomedical data to aid the diagnostic process by understanding and discussing these visuals through a natural language dialogue with clinicians.

4.3. LLaMA

The LLaMA (Large Language Model Meta AI) collection, was unveiled in 2023 by Touvron et al. [68]. It brings together a whole series of foundational models of imposing size, varying from 7 to 70 billion parameters. These giant models could be trained on a massive scale via unsupervised learning techniques such as word masking and next-sentence prediction. The training was done on colossal public datasets, representing trillions of tokens in total.

These resources included Wikipedia, Common Crawl, and OpenWebText. By using only freely accessible data, the creators of LLaMA have shown that it is possible to achieve a state-of-the-art level of performance without resorting to proprietary datasets. This collection thus demonstrates the remarkable progress made in the field, now allowing the deployment of giant linguistic models trained on a very large scale on public resources.

The LLaMA-1 model [68] was developed with efficient causal attention [102] by avoiding storing and calculating masked attention weights and key/query scores. A further optimization was achieved by reducing the number of activations recalculated during backtracking, as described in [103]. As part of the work on LLaMA-2 [69], the focus was on improving a specific model called LLaMA-2-Chat, making it safer and more efficient for dialogue generation. The LLaMA-2 pre-trained model has been improved by incorporating 40% more training data. In addition, the context length was extended and a feature called Grouped Query Attention (GQA) was added [104]. The model was trained on a massive dataset consisting of 2000 billion tokens.

Gema et al. [105] have introduced Clinical LLaMA-LoRA, a Parameter-Efficient Fine-Tuning (PEFT) adaptation layer based on the LLaMA model. This specific adaptation for the clinical domain is trained using clinical notes from the MIMIC-IV database. By combining these data, a specialized adapter is created to improve the model's performance in the clinical domain. In addition, the authors propose a two-stage PEFT framework that integrates both Clinical LLaMA-LoRA and Downstream LLaMA-LoRA. The latter PEFT adapter is specifically designed for downstream tasks, enabling better adaptation of the model to specific tasks in the clinical domain.

PMC-LLaMA [106] is an open-source language model specially designed for the analysis of biomedical articles. It is fine-tuned from LLaMA on biomedical academic articles, giving it increased specialization and understanding of biomedical language.

Visual Med-Alpaca [107] is a multimodal biomedical model based on the LLaMa-7B architecture. It is specially designed to handle various biomedical tasks by integrating both linguistic and visual data. Its training process involves a collaboration between GPT-3.5-Turbo, a powerful language model, and human experts.

4.4. Bloom

Bloom is a massive multilingual LLM with 176 billion parameters, trained on the NVIDIA AI platform using vast text data [89].

While referred to as an LLM, Bloom's core functionality is in text generation through continuation; it is prompted with an initial context and asked to complete and extend the text. The terms generation, continuation, and completion are thus used somewhat interchangeably to describe Bloom's text production abilities.

A key attribute of Bloom is its multilingualism; it can generate text in 46 different human languages as well as 13 programming languages. By leveraging immense computational resources, Bloom was trained on an industrial scale with massive language datasets, resulting in a highly capable generative model with a breadth of linguistic competencies unprecedented for an openly available LLM.

As described in [89], BLOOM's architecture and pre-training are based on the transformer approach, using a causal-only decoder model. BLOOM's modeling details include the use of ALiBi positional embeddings, which directly adjust attention scores according to the distance between keys and queries. This facilitates training and improves performance. In addition, an embedding normalization layer is added after the integration layer to stabilize training, using float16 precision.

In addition to these architectural components, data pre-processing is crucial. This includes steps such as de-duplication and privacy suppression, particularly for high-risk sources. BLOOM also employs prompted fine-tuning with guest datasets, which enhances its zero-shot generalization capabilities and improves performance.

4.5. StableLM

StableLM, an open-source LLM, has been launched by Stability AI [91]. It is available in versions with 3 billion and 7 billion parameters, with larger versions on the horizon (with 15 billion to 65 billion parameters). This follows the introduction in 2022 of Stable Diffusion, a model for generating images from a sentence of text. StableLM, designed to generate text and code, shows that smaller models can achieve high performance with the right training. The model relies on a large dataset based on the Pile dataset, but three times as large. This large dataset enables the model to excel in conversational and coding tasks, even though its number of parameters is considerably lower than that of models such as GPT-3.

StableLM comes in two main variants: StableLM-3B-4E1T, which focuses on studying the impact of repeated tokens, and StableLM-Alpha v2, which improves on the initial architecture of the Alpha models and uses larger, higher-quality datasets. In addition, experimental fine-tuning has been carried out to develop StableLM-Tuned-Alpha, combining various datasets to enhance its capabilities as a conversational agent.

5. Applications of Transformer-Based LLMs in Healthcare

When a patient meets a clinician for the first time, it is essential to establish an accurate diagnosis and create a relationship of trust. Unfortunately, time constraints often limit the ability to thoroughly review medical histories and provide tailored care to each patient. LLMs can improve this process by extracting relevant information from electronic health records [108], allowing a concise overview of the patient's medical history to be presented. This helps optimize consultation productivity.

By highlighting past medical conditions, treatments, medications, and results of previous visits, LLMs eliminate the need to review numerous records, allowing for more targeted interactions to address specific patient issues. By focusing on key concerns, LLMs ensure that diagnosis is personalized to each patient. Additionally, these models are constantly improved as health data grow, ensuring their accuracy. By ensuring data transparency and implementing continuous monitoring, the potential of LLMs can be further improved, resulting in faster diagnoses and greater patient satisfaction.

Transformer-based language models, such as LLaMA, ChatGPT [109], and GPT-4 [2] have shown great potential in a variety of fields, including medical practice. These models are particularly efficient at understanding and producing texts that resemble those written by humans, opening up possibilities for supporting healthcare professionals in the diagnostic process, as shown in Figure 5. The rest of this section illustrates how these models can be applied for this purpose.

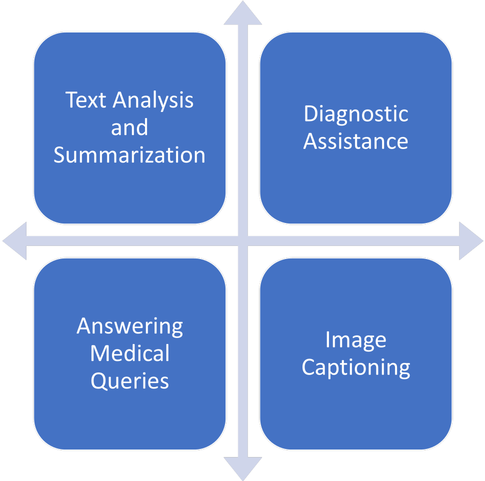

Figure 5. Applications of transformer-based LLMs in healthcare.

5.1. Text Analysis and Summarization

Electronic medical records (EMRs) have transformed the way healthcare professionals access and use patient data [110], profoundly altering their approach to decision-making. A major advantage of EMRs lies in their ability to summarize clinical observations [18], giving doctors a clear view of potential health hazards and facilitating informed medical choices. This data summarization not only minimizes diagnostic errors but also optimizes therapeutic outcomes thanks to up-to-date, targeted patient information [111].

Nevertheless, manual summarization of clinical reports is laborious and error-prone [112]. The abundance and density of data mean that even seasoned professionals can miss crucial details. It is therefore imperative to develop automated synthesis techniques to increase the efficiency and reliability of medical care.

Relying on these automated summarization techniques, medical staff could rapidly extract the essential elements from the clinical notes provided, thereby reducing the risk of errors and enhancing the quality of the care provided.

The rise of LLMs in NLP has opened new possibilities for streamlining automated clinical report summarization. These models, recognized for their capability to align with input directions, benefit from the integration of text prompts [113,114].

Indeed, LLMs, like BERT, BART, LLAMA, Bloom, GPT-3.5, and GPT-4, have demonstrated a remarkable ability to analyze, understand, and summarize large quantities of unstructured text such as clinical notes, patient histories, scientific articles, etc. [115,116]. This makes them helpful in extracting meaningful information from textual medical data [117].

Critical applications include automating the extraction of crucial information from detailed medical records [118–121], creating customized summaries to assist healthcare professionals [122,123], supporting report-based coding and billing [123,124], and semantically grouping symptoms to facilitate diagnostic processes [125].

Large amounts of medical literature can also be filtered and analyzed by LLMs, which can then dynamically summarize data from many sources for doctors [126,127]. Succinctly synthesizing ideas saves crucial time.

LLMs can analyze clinical text and identify clinical concepts, entities, and their interrelationships. This enables them to generate differential diagnoses [128,129], predict specific diagnoses [130,131], and provide answers to physicians' queries [98,132]. Moreover, LLMs' multilingual capabilities can facilitate global healthcare [133].

In capturing patient details and symptoms, LLMs offer a promising solution for the extensive documentation that physicians and clinical experts face daily [134]. These models can produce detailed clinical summaries and diagnostic reports, effectively alleviating the time-intensive burden of these professionals [135]. A concrete example of their potential is an LLM specially designed to condense the results of radiology reports, highlighting its applicability in similar medical fields [136].

To optimize time management during consultations, LLMs can be exploited to generate concise summaries of each patient's medical history. These summaries include information on comorbidities, previous consultations or admissions, medication lists, as well as progress and response to previous treatments [19,137]. These synthesized summaries draw the doctor's attention to relevant information about the patient, which is crucial for an accurate diagnosis of the disease. By concisely providing this information, LLMs promote more efficient and comprehensive consultations. This approach allows doctors to devote more time to patient interaction, which in turn increases patient satisfaction.

Joseph et al. [138] recently introduced FACTPICO, an evidence-based reference for plain-language summaries of medical texts describing randomized controlled trials (RCTs). RCTs are essential in evidence-based medicine and play a direct role in patients' treatment decisions. FACTPICO consists of 345 plain-language abstracts generated by three LLMs (GPT-4, Llama-2, and Alpaca). It includes fine-grained evaluation and natural language justifications provided by experts.

As a result, using LLMs to generate concise summaries optimizes time management during consultations, enables physicians to focus on patient interaction, and improves patient care [139]. In this way, the summary process helps clinicians to efficiently review patient histories, research studies, and guidelines [140].

5.2. Diagnostic Assistance

Using LLMs to aid medical professionals in making diagnoses has shown potential [141]. LLMs, trained on a massive medical corpus, have acquired important clinical knowledge. This latent expertise gives them several assets that can support the diagnostic process.

LLMs can synthesize information from a variety of sources, such as medical literature, clinical guidelines, and case histories, to guide decision-making. Using a patient's data, these models can automatically generate differential diagnosis hypotheses, ranked according to their estimated probability, for consideration by the clinician [142–144].

Medical doctors (MDs) can quickly acquire pertinent decision assistance by using natural language queries to interact with LLMs through conversational interfaces. In order to identify patterns and create clinical syndromes, related indications and symptoms may be semantically grouped [145,146].

LLMs can also assist in clinical decision-making by suggesting suitable medications [147], recommending relevant imaging services based on symptoms [148], or pinpointing disease causes from diverse clinical documents. When combined with tools like medical imaging, these models can offer a holistic view, helping doctors diagnose and define diseases [149]. Moreover, by evaluating cases of individuals with comparable symptoms, LLMs can forecast potential disease outcomes, empowering both doctors and patients to make well-informed treatment choices.

By extracting the most important elements from narrative records, LLMs can simplify the review process for healthcare professionals by creating structured record segments [150,151]. In addition, the synthesis of information from exchanges or expert research helps to improve data assimilation.

Supplementary examinations often play a crucial role in medical diagnosis, helping to reinforce clinical hypotheses and confirm diagnoses. In this context, LLMs can act as clinical decision support tools, helping physicians choose the most relevant radiological examinations for given clinical cases [8]. This approach has the potential to minimize pressure on limited resources, particularly in public hospitals or resource-constrained areas where imaging equipment and technical support are limited. Furthermore, LLMs can help avoid unnecessary examinations, such as those involving radiation or contrast agents, for patients who do not need them.

The increasing use of DL models to create computer-aided diagnosis (CAD) systems has automated the interpretation of medical images, facilitating the detection and classification of pathologies. However, their integration into clinical practice is sometimes hampered by a lack of transparency in the decisions made by these models. To remedy this, the integration of LLMs into CAD systems can enable clinicians to ask questions about specific images or patient cases, thereby clarifying the CAD system's decision-making process. This human–machine interaction can make models more interpretable and encourage their use in routine diagnostic procedures. In addition, this approach may reveal new insights or biomarkers in disease imaging.

A recent study [31] presents a new paradigm called generalist medical AI (GMAI), in which models are trained on large, diversified medical datasets. This approach allows models to perform a large range of tasks, such as diagnosis, prognosis, and treatment planning. The paper evaluates the performance of GMAI models on different medical tasks and finds that they perform better than conventional specialized medical AI models in several areas, including diagnosis, prognosis, and treatment planning.

In the future, the integration of different data sources, such as images and laboratory tests, could provide more comprehensive diagnostic information [152]. Throughout this process, however, it is essential to maintain the interpretability of models and to assign responsibility for results to human practitioners. The integration of multimodal data will pave the way for in-depth diagnostic analysis [153].

Despite the significant progress made by LLMs, their use to assist medical diagnosis has several important limitations. First, these models may lack the clinical precision necessary for specific diagnoses, as they may fail to capture all essential details or misinterpret complex medical information. Additionally, they generally struggle to fully understand the clinical context of a patient, including crucial elements such as medical history or current symptoms [154].

The data used to train these models can also contain biases or inaccuracies, which can translate into erroneous recommendations. It is important to emphasize that LLMs can in no way replace human clinical expertise, especially regarding physical examination and interpretation of medical test results.

Data privacy and security issues are also a major concern, particularly when dealing with sensitive patient information. Furthermore, LLMs are not always effective in interpreting complex medical test results, such as radiological images or laboratory tests. Their use in the medical field therefore requires rigorous clinical validation to ensure their safety and efficacy before deployment in a healthcare setting.

Finally, given the rapidly evolving nature of the medical field, these models need to be constantly updated to incorporate new discoveries and practices, which necessitates continuous maintenance.

5.3. Answering Medical Queries

QA is a task that automatically provides an answer to a given question. Indeed, LLMs have greatly progressed the field of QA in NLP, enabling machines to comprehend and respond to natural language queries with precision. These extensive language models find applications in virtual assistants and chatbots, delivering precise and pertinent responses to user questions. Such systems leverage LLMs to grasp the context and meaning of inquiries and formulate suitable replies. For instance, Google Assistant, underpinned by LLMs, adeptly addresses an array of user queries spanning general knowledge, weather updates, directions, and more [57].

Advancements in DL exemplified by "transformers" now open up new possibilities [53]. Under their enhanced capacities, LLMs [4,55] have the potential to gain a finer-grained understanding of complex relationships within enormous corpora. The recent emergence of transformers and LLMs has breathed new life into research exploring AI's potential to tackle the great challenge [155,156] of medical QA.

Previously, most works relied on smaller linguistic models trained specifically on domain-specific healthcare text corpora [94,157–159]. This has driven steady improvements in benchmark performance on reference tests such as MedQA [160], MEDMCQA [161], SentiMedQAer [162], and DATLMedQA [163].

Answering medical queries using LLMs has shown significant advancements in recent times. With the emergence of larger general-purpose LLMs like GPT-3 [3], GPT-Neo [164], GPT-3.5 [1], OPT [165], LLaMA [68], and Flan-PaLM [84,87], trained on massive internet-scale datasets using extensive computational resources, there has been a remarkable improvement in their performance on medical QA benchmarks.

For instance, GPT-3.5 achieved an accuracy of 60.2% on the MedQA (USMLE) dataset, demonstrating its ability to provide accurate responses to medical queries [98,166]. Flan-PaLM, another powerful LLM, achieved an even higher accuracy of 67.6% on the same dataset. The performance of GPT-4-base [2] was even more impressive, achieving an accuracy of 86.1% on medical QA tasks [167,168].

These advancements in LLMs have been observed within a relatively short time, showcasing their rapid progress and potential for accurate and reliable medical QA. It is important to note that these accuracy figures are specific to the mentioned datasets and models, and the performance may vary depending on the specific task and dataset involved.

The critical question remains whether they can generate responses meeting the demands of the clinical context—reliability, interpretability, and adherence to ethical standards, prerequisites for actual utility at the point of care [169–172].

One of the major challenges encountered in medical QA systems is the problem of hallucinations leading to incorrect answers [173]. An approach commonly used to solve this problem is retrieval augmentation. This approach consists of combining LLMs with a search system such as New Bing for general domains, or Almanac for clinical domains [174]. It involves first retrieving relevant documents as support, and then using LLMs to answer a specific question based on these retrieved documents. The underlying idea is that LLMs are able to effectively summarize content, which can reduce hallucinations. It is important to stress, however, that these systems are not error-free [175] and require thorough and systematic evaluation to ensure their quality [176]. Further studies are needed to rigorously evaluate the performance of these systems and identify possible sources of error.

A promising approach to solving the hallucination problem is to augment LLMs with additional tools. For example, recent work has explored the use of LLM augmentation by combining data from specific sources [177–180]. A concrete example is the GeneTuring dataset, which contains search questions for information on specific single nucleotide polymorphisms (SNPs). However, it is important to note that autoregressive LLMs do not possess specific knowledge about these SNPs, and most commercial search engines are unable to provide relevant results for such queries. Consequently, the approach of increasing retrieval may prove ineffective in such cases.

In such situations, relevant information is often only accessible via specialized databases such as NCBI dbSNP. For this reason, the approach of enriching LLMs using web database utility APIs, such as those provided by NCBI, offers the potential for solving the problem of hallucinations related to specific entities in biomedical databases [178].

While computational progress hints at promising avenues, rigorous validation in real-world healthcare environments remains essential before these systems can meaningfully impact patient outcomes [176]. Overall, advancements in NLP promise to improve the ability of systems to answer medical questions. However, this is only possible if rigorous evaluations are carried out to ensure the safety, transparency, and responsibility of such solutions. Indeed, as healthcare is a crucial area where the smallest failure could have serious consequences, it is essential to ensure through thorough testing that digital medical assistance systems respect the highest standards in terms of ethics, protection of sensitive data, and patient well-being.

5.4. Image Captioning

Developing precise and dependable automated report-generation systems for medical images presents various obstacles. These include the analysis of limited medical image datasets using machine learning techniques and the generation of informative captions for images that involve multiple organs.

The generation of image captions has been the subject of considerable work in computer vision. Early models focused on characterizing the objects, attributes, and relations present in the image via visual primitive extraction methods [181–184].

The emergence of DL enabled the development of end-to-end neural architectures that encode an image as a vector representation and then generate the legend word by word [185,186]. Since then, multiple improvements have been made to encoding [187–192] and decoding [193–195] blocks, as well as to attention mechanisms [196–199].

It has been shown that encoding by object region instead of global image representation increases performance [200]. These advancements in image encoders and decoders, together with the development of attention networks, have led to significant improvements in the quality of automatically generated captions.

Captioning medical images is an especially difficult task due to the complexity and variability of images such as fetal ultrasound images. The latter are usually noisy, of low resolution, and dependent on factors such as fetal position, gestational age, or imaging plan. Furthermore, the availability of extensive annotated datasets is essential for tackling this task effectively. In response to these challenges, researchers have put forward DL-based approaches that seamlessly merge visual and textual data, enabling the creation of informative captions for both images and videos.

Alsharid et al. [201] presented a model for image captioning specifically designed for fetal ultrasound images. Alsharid et al. [202,203] have taken their studies further by introducing a programmatic learning approach to train such image captioning models. They have also proposed a captioning framework where an image is first classified, and then one of the multiple-image captioning models is employed to generate the corresponding caption. Each image captioning model is associated with an anatomical structure.

In a separate investigation [204], researchers addressed the difficulties encountered in generating captions for medical images, particularly those involving multiple organs. In response, they propose a solution that combines DL and transfer learning techniques, specifically employing a Multilevel Transfer Learning Technique (MLTL) and LSTM framework. The authors introduce a foundational MLTL framework consisting of three models designed to detect and classify datasets with severe limitations. This is achieved by leveraging knowledge gained from readily available datasets. The first model utilizes non-medical images to acquire generalized features, which are subsequently transferred to the second model, responsible for the intermediate and auxiliary domains related to the target domain. The study concludes by discussing the potential applications of this approach in the medical diagnosis field and its potential to enhance patient care.

A new model for automatic clinical image caption generation has been developed [205]. It combines radiological scan analysis with structured patient information to generate comprehensive and detailed radiology reports. The model utilizes two language models, Show-Attend-Tell [206] and GPT-3, to generate captions that contain important information about identified pathologies, their locations, and 2D heatmaps highlighting the pathologies on the scans. This approach improves the interpretation and communication of radiological results, facilitating clinical decision-making.

The application of caption generation is not limited solely to images; it extends to videos as well. As part of another research project [207], the authors introduced an innovative method for generating captions for fetal ultrasound videos, which can be valuable in aiding healthcare professionals in their diagnosis and treatment decision-making processes. This approach leverages a three-way multimodal deep neural network that merges gaze-tracking technology with NLP. The primary objective is to develop a comprehensive understanding of ultrasound images, enabling the generation of detailed captions encompassing nouns, verbs, and adjectives. The potential applications of these DL models include integration into systems designed to assist in the interpretation of ultrasound scan video frames and clips. The article also explores previous studies in the field of image and video captioning and presents quantitative assessment findings.

Li et al. [208] developed a novel end-to-end video understanding system called VideoChat, which focuses on chat interaction. The system incorporates video foundation models and LLMs via a learnable neural interface. Its main feature is its ability to perform spatiotemporal reasoning, locate events in the video, and infer causal relationships between them. To fine-tune the system, the researchers developed a specific dataset focused on video instructions. This comprehensive dataset includes thousands of videos with detailed descriptions and associated conversations. It places particular emphasis on spatiotemporal reasoning and captures the causal relationships between events in the videos.

Although important advancements have been made in the field of generating captions for medical images and videos using LLMs, certain challenges still need to be tackled. Firstly, the limited size of the datasets used in some studies represents one of the main difficulties, as it may compromise the generalizability of the proposed approaches to more expansive datasets. In addition, the evaluation measures used in some research studies do not always provide a comprehensive understanding of the quality of the generated captions. In the future, it would be interesting to explore more sophisticated measures that would enable us to evaluate model performance more effectively. Improving the representativity of the data used and refining the evaluation criteria would be promising avenues for progress in this field.

Another challenge lies in the need for large quantities of annotated data for model training, which can prove complicated in the field of medical imaging where annotation requires expertise. In the future, research could explore new ways of reducing dependence on massive quantities of pre-annotated data, for example by developing methods that enable better generalization of the proposed approaches. The integration of additional contextual information such as clinical metadata or patient medical history into the caption generation process also represents a promising avenue for providing richer context and improving both the relevance and accuracy of the captions produced. This type of multimodal approach would go some way to overcoming the lack of massive annotated data.

Future research in medical image and video captioning can focus on two key areas: advancing transformer-based word embedding models and spatiotemporal visual and textual feature extractors and utilizing larger and more diverse datasets. These advancements aim to improve the accuracy, precision, and usefulness of generated captions in medical applications. By developing more sophisticated models and incorporating richer datasets, the goal is to overcome limitations such as dataset size and lack of diversity, ultimately enhancing the performance of automatic report generation systems in the medical field.

6. Analysis of Massive Medical Datasets

In this section, as shown in Figure 6, we will present a range of datasets by classifying them into three categories: general datasets, de-identification datasets, and medical QA datasets.

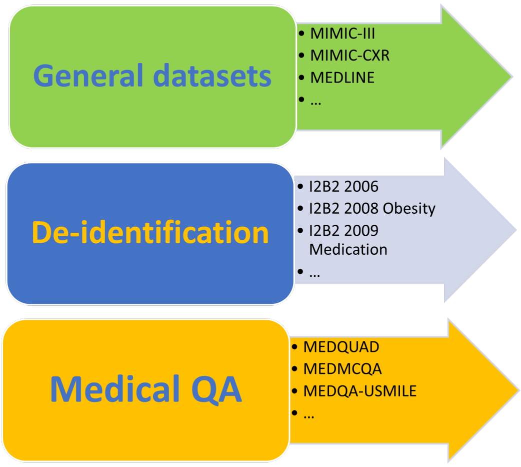

Figure 6. Medical Datasets

6.1. General Datasets

In this section, we present a non-exhaustive overview of several commonly used medical datasets, with a synthetic summary of their main characteristics, as described in Table 3.

This table succinctly summarizes key information on various datasets representative of clinical and biomedical fields, such as their theme, approximate size, and type of content, as well as a brief description of some.

Although this list is not intended to be exhaustive, it does provide an overview of the resources available and frequently exploited in current research in automatic medical language processing.

Table 3. Summary of general medical datasets.

Dataset	Domain	Size	Data Type
MIMIC-III	Intensive care	Over 58,000 patients	Structured and unstructured clinical data
MIMIC-CXR	Radiology	Over 380,000 images	Medical images (X-rays)
MEDLINE	Biomedical	Over 25 million abstracts	Biomedical article abstracts
ABBREV	Biomedical	6205 abbreviations	Abbreviation list
PUBMED	Biomedical	Over 30 million abstracts	Article abstracts
NUBES	Spanish biomedical	30,000 phrases	Phrases
NCBI	Disease	11,223 documents	Texts
CASI	Terminology	6300 entries	Abbreviation senses
MEDNLI	Clinical reasoning	10,000 sentence pairs	Sentence pairs
MEDICAT	Clinical trials	30,000 trials	Semi-structured clinical trial descriptions
OPEN-I	Adverse drug reactions	Over 1 million ADRs	Doctor narratives about ADRs
MEDICAL ABSTRACTS	Biomedical	Over 400,000 abstracts	Scientific paper abstracts

1. *MIMIC-III*, known as "Medical Information Mart for Intensive Care III", is a publicly accessible database focused on critical care [209]. It is an extensive and comprehensive health database that includes de-identified health-related data from more than 40,000 patients who were admitted to Beth Israel Deaconess Medical Center (BIDMC) in Boston between the years 2001 and 2012. The database encompasses a diverse array of information, including clinical notes, physiological waveforms, laboratory test results, medication details, procedures, diagnoses, and demographics. This extensive dataset holds immense value for medical research and healthcare analytics, as well as the creation and validation of machine learning models and clinical decision support systems. It provides a valuable resource for advancing medical knowledge and enhancing patient care. MIMIC-III is widely utilized by researchers and healthcare professionals for a range of studies, including predictive modeling [210,211], risk stratification [212], treatment outcomes analysis [213], and other medical research investigations [214,215]. The database offers valuable insights into patient care [216,217], facilitates the development of advanced healthcare technologies [218,219], and contributes to the enhancement of clinical practices and patient outcomes [220,221]. It is essential to emphasize the importance of ethical considerations and strict adherence to data usage policies when accessing and utilizing the MIMIC-III database to ensure proper handling and protection of patient information.
2. *MIMIC-CXR*, referred to as "Medical Information Mart for Intensive Care—Chest X-ray", is an expansion of the MIMIC-III database that specifically concentrates on chest X-ray images and related clinical data [222]. This publicly available dataset comprises de-identified chest X-ray images alongside their corresponding radiology reports for a substantial number of patients. A total of 377,110 images are available in the MIMIC-CXR dataset. These images are associated with 227,835 radiographic stud-

ies conducted at the Beth Israel Deaconess Medical Center in Boston. The contributors to the dataset utilized the ChexPert tool [223] to classify the free-text notes associated with each image into 14 different labels. This process categorized the textual information accompanying the images, providing additional context and information for research and analysis. MIMIC-CXR is a valuable dataset that plays a crucial role in training and evaluating machine learning models and algorithms focused on chest X-ray image analysis, radiology report processing, NLP, and various medical imaging tasks. Researchers and healthcare professionals rely on MIMIC-CXR to develop and validate AI-driven systems designed for automated diagnosis [224,225], disease detection [226,227], and image-captioning applications specifically tailored to chest X-ray images [205,228].

3. *MEDLINE* is an extensive and highly regarded bibliographic database that encompasses life sciences and biomedical literature. It is curated and maintained by the National Library of Medicine (NLM) and is a component of the larger PUBMED system. MEDLINE contains more than 29 million references sourced from numerous academic journals, covering a wide range of disciplines including medicine, nursing, dentistry, veterinary medicine, healthcare systems, and preclinical sciences. MEDLINE is a widely utilized resource by researchers, healthcare professionals, and scientists seeking access to an extensive collection of scholarly articles and abstracts. It serves as a vital source of information for academic research, aiding in clinical decision-making [229,230], and keeping individuals informed about the latest developments in the medical and life sciences [231–233]. The database plays a crucial role in supporting evidence-based practice, enabling professionals to stay updated with the most recent advancements and discoveries in their respective fields. MEDLINE's comprehensive coverage and wealth of information make it an essential tool for professionals and researchers across the medical and life sciences domains. In MEDLINE, the data commonly consist of bibliographic information such as article titles, author names, abstracts, publication sources, publication dates, and other pertinent details. This dataset is an essential and foundational resource for conducting research in the fields of medicine and life sciences, supporting diverse applications like NLP [234], information retrieval [235,236], data analysis [237], and more [238,239]. The wealth of information contained within MEDLINE enables researchers to explore and extract valuable insights, contributing to advancements in medical knowledge and facilitating a wide range of research endeavors within the healthcare and life sciences domains [240–242].
4. *ABBREV dataset*, proposed by Stevenson et al. in 2009 [243], is a collection of acronyms and their corresponding long forms extracted from MEDLINE abstracts. Originally introduced by Liu et al. in 2001, this dataset has undergone automated reconstruction. The reconstruction process involves identifying the long forms of acronyms in MEDLINE and replacing them with their respective acronyms. The dataset is divided into three subsets, with each subset containing 100, 200, and 300 instances, respectively.
5. *The PUBMED* dataset comprises over 36 million citations and abstracts of biomedical literature; while it does not provide full-text journal articles, it typically includes links to the complete texts when they are available from external sources. Maintained by the National Center for Biotechnology Information (NCBI), PUBMED is an openly accessible resource for the public. Serving as a comprehensive search engine, it enables users to explore a vast collection of articles from diverse biomedical and life science journals. PUBMED encompasses a wide range of subjects, spanning medicine, nursing, dentistry, veterinary medicine, biology, biochemistry, and various other fields within the biomedical domain. Within the PUBMED dataset, you can find a wealth of information, such as article titles, author names, abstracts, publication sources, publication dates, keywords, and MeSH terms (Medical Subject Headings). This extensive dataset is extensively used by researchers, healthcare professionals, and individuals within the academic and medical communities. PUBMED serves as a

go-to platform for accessing the latest research and information in the vast field of biomedicine. It provides a comprehensive search engine that enables users to explore a diverse range of topics, including medicine, nursing, dentistry, veterinary medicine, biology, biochemistry, and more. By utilizing PUBMED, researchers and professionals can stay updated with the latest scientific literature, conduct literature reviews, and make evidence-based decisions in their respective fields.

6. *The NUBES dataset* [244], short for "Negation and Uncertainty annotations in Biomedical texts in Spanish", is a collection of sentences extracted from de-identified health records. These sentences are annotated to identify and mark instances of negation and uncertainty phenomena. To the best of our knowledge, the NUBES corpus is presently the most extensive publicly accessible dataset for studying negation in the Spanish language. It is notable for being the first corpus to include annotations for speculation cues, scopes, and events alongside negation.
7. *The NCBI disease corpus* comprises a massive compilation of biomedical abstracts curated to facilitate NLP of literature concerning human pathologies [245]. Leveraging Medical Subject Headings assigned to articles within PUBMED, the corpus was constructed by annotating over 1.5 million abstracts that discuss one or more diseases according to controlled vocabulary terms. Spanning publications from 1953 to the present day, the breadth of the included abstracts encapsulates a wide spectrum of health conditions. Each is tagged with the relevant disease(s) addressed, permitting focused retrieval. Beyond these annotations, access is also provided to the original PUBMED records and associated metadata, such as publication dates. By manually linking this substantial body of literature excerpts to standardized disease descriptors, the corpus establishes an invaluable resource for information extraction, classification, and QA applications centered around exploring and interpreting biomedical writings covering an immense range of human ailments. The colossal scale and selection of peer-reviewed reports ensure the corpus offers deep pools of content for natural language models addressing real-world biomedical queries or aiding disease investigations through text-based analysis. Overall, it presents a comprehensive, expertly curated foundation for driving advancements in medical text mining.
8. *The CASI (Clinical Abbreviation Sense Inventory) dataset* aims to facilitate the disambiguation of medical terminology through the provision of contextual guidance for commonly abbreviated terms. Specifically, it features 440 of the most prevalent abbreviations and acronyms uncovered among the 352,267 dictated clinical notes. For each entry, potential intended definitions, or "senses", are enumerated based on an analysis of the medical contexts within which the shorthand appeared across the vast note compilation. By aggregating examples of appropriate abbreviation implementation in genuine patient encounters, the inventory helps correlate ambiguous clinical jargon with probable denotations. This aids in navigating the challenges inherent to interpreting abbreviated nomenclature pervasive throughout medical documentation, which is often polysemous. The dataset provides utility for natural language systems seeking to map abbreviations to intended clinical senses when processing narratives originating from authentic healthcare records.
9. *The MEDNLI dataset* was conceived with the goal of advancing natural language inference (NLI) within clinical settings. As in NLI generally, the objective involves predicting the evidential relationship between a hypothesis and premise as true, false, or undetermined. To facilitate model progress, MEDNLI comprises a sizable corpus of 14,049 manually annotated sentence pairs. The data originate from MIMIC-III, necessitating initial access to extract the pairs, while retrieval is dependent on securing permission to access the source's protected patients, MEDNLI balances this ethically with the availability of an extensive set. This supports valuable work on a key clinical NLP challenge: determining the inferential relationship between ideas expressed.
10. *The MEDICAT dataset* comprises a vast repository of medical imaging data, matched materials, and granular annotations [246]. It contains over 217,000 images extracted

from 131,410 open-access articles in PUBMED Central. Each image is accompanied by a caption and 7507 feature compound structures with delineated subfigures and subcaptions. The reference text comes from the S2ORC database. The collection also includes online citations for around 25,000 images in the ROCO subset. MEDICAT introduces a new method for annotating the constituent elements of complex images. It correlates over 2000 composite visualizations with their constituent subfigures and descriptive subcapitals. This refined partitioning at multiple intralimatic scales within individual elements far surpasses previous collections based on unitary image-caption couplings. MedICaT's granular annotations lay the foundations for modeling intra-element semantic congruence, an essential capability for tackling the sophisticated challenges of visual language in biomedical imaging. With its vast scale and careful delineation of image substructures, MEDICAT provides a cutting-edge dataset to advance research at the intersection of vision, language, and clinical media understanding.

11. *The OPEN-I, the Indiana University Chest X-ray dataset*, presents a sizable corpus of 7470 chest radiograph images paired with associated radiology reports [247]. Each clinical dictation encompasses discrete segments for impression, findings, tags, comparison, and indication. Notably, rather than using full reports as captions, this dataset strategically concatenates the impression and findings portions for each image, while impressions convey overarching diagnoses, findings provide specifics on visualized abnormalities and lesions. By combining these sections, the designated captions offer concise and comprehensive clinical overviews. With a sample size exceeding 7000 image–report duos, this collection equips researchers with ample training and evaluation materials. Distinct from simpler descriptors, the report snippets incorporated as captions encompass richer diagnostic particulars. This notation format furnishes target descriptions well-suited to nurturing technologies aimed at automating radiographic report generation directly from chest X-ray visuals.
12. *The MEDICAL ABSTRACTS dataset* presents a robust corpus of 14,438 clinical case abstracts describing five categories of patient conditions, an integral resource for the supervised classification of biomedical language [248]. Unlike many datasets, each sample benefits from careful human annotation, providing researchers with a fully labeled collection that avoids ambiguity. The abstracts also emanate from real medical scenarios, imbuing the models with authentic representations. The collection divides examples into standardized training and test scores, enabling a rigorous and consistent evaluation of classification performance. At this scale, DL techniques can derive profound meaning from the distribution of terminology across different specialties. Automatic organization of vast quantities of documents according to disease status should lead to improved indexing, keyword assignment, and file routing. This fundamental resource enables investigators to cultivate methods that intelligently classify and route written summaries, streamlining discovery and care.

6.2. De-Identification Dataset

I2B2 (Informatics for Integrating Biology and the Bedside) is a non-profit entity with the primary goal of organizing NLP competitions and assembling datasets centered around clinical language. These datasets are tailored for specific tasks like de-identification, extracting relationships, and selecting cohorts for clinical trials. Despite the management transition to Harvard's National Clinical Language Challenges (N2C2), academic research predominantly continues to cite these datasets under the I2B2 name, acknowledging their founding influence and enduring significance in the domain, as shown in Table 4. Hereafter, we will detail all datasets associated with I2B2:

Table 4. Summary of de-identification datasets.

Dataset	Domain	Size	Data Type
I2B2 2006	De-identification	347 documents	Medical records
I2B2 2008 Obesity	Phenotyping	571 documents	Medical records focused on obesity
I2B2 2009 Medication	Relation extraction	402 documents	Medical records
I2B2 2010 Relations	Relation extraction	947 documents	Medical records
I2B2 2011 coreference	Coreference resolution	970 documents	Medical records
I2B2 2012 Temporal	Temporal relation extraction	2000 medical encounters	Medical records
I2B2 2014 Deidentification	Heart Disease dataset	Focused on cardiovascular diseases	Medical records
I2B2 2018	Benchmarking	Hierarchical clinical document classification	Clinical case notes

1. *The I2B2 2006 dataset* serves two main objectives—deidentification of protected health information from medical records [249], and extraction of smoking-related data [250]. Regarding privacy, it focuses on removing personally identifiable information (PII) from clinical documents. This is important for enabling the ethical and lawful use of such records for research purposes by preserving patient anonymity. At the same time, the dataset facilitates the extraction of key smoking-related details from these records. This aspect is crucial for understanding impactful health behavior and risk factors. Researchers leverage this dual-purpose dataset to develop automated systems capable of both tasks through NLP. This significantly advances the application of NLP within healthcare to tackle real-world issues surrounding privacy and extracting meaningful insights from unstructured notes. The dataset is freely available to the research community. Overall, it serves an important role in propelling research at the intersection of privacy, smoking cessation, and NLP technologies for EMR.
2. *The I2B2 2008 Obesity dataset* is a specialized collection aimed at facilitating the recognition and extraction of obesity-related information from medical records [251]. The dataset challenge centered on identifying various aspects of obesity within clinical notes, such as body mass index, weight, diet, physical activity levels, and related conditions. Researchers leverage this curated corpus to build models and algorithms capable of accurately detecting and extracting obesity-related data from unstructured notes. Key elements identified include BMI, weight measurements, dietary patterns, exercise habits, and associated medical conditions. By developing technologies to systematically organize this clinical information, researchers can gain deeper insights into obesity trends, risk factors, and outcomes. This contribution ultimately advances healthcare research and strategies for obesity treatment and prevention [252–254]. The specificity of the I2B2 2008 Obesity dataset in regards to this important public health issue has made it a valuable resource for the medical informatics community to progress solutions.
3. *The I2B2 2009 Medication dataset* is a specialized corpus focused on extracting granular medication-related information from clinical notes [255]. The dataset challenge centered on precisely identifying and categorizing various facets of medications documented within notes, such as names, dosages, frequencies, administration routes, and intended uses. The corpus is meticulously annotated to demarcate and classify mentions of medications and associated attributes in the unstructured text [256]. Researchers leverage this resource to develop sophisticated NLP models tuned to accurately recognize and extract intricate medication details. Key elements identified and disambiguated include drug names, dosage amounts, dosing schedules, administration methods, and therapeutic purposes. By automatically organizing these clinical aspects at scale, the resulting NLP systems enable vital insights to support health-

care decision-making and research [257]. The specificity and detailed annotation of the dataset in mining this significant clinical domain has made it instrumental for advancing automatic extraction of medication information [258,259].

4. *The I2B2 2010 Relations dataset* was specially curated to facilitate extracting and comprehending associations between medical entities in clinical texts. The dataset challenge centered on accurately identifying and categorizing relationship types expressed in clinical narratives, such as drug–drug interactions or dosage frequency links [260]. Annotations clearly define and classify the annotated relations, supporting the development of models precisely capable of recognizing and categorizing diverse medical connections. Advancements ultimately aim to bolster comprehension of treatment considerations and implications to elevate standards of care based on a complete view of patient context and interconnectivity [261,262].
5. *The I2B2 2011 coreference dataset* has been specially designed to facilitate the resolution of coreferences in clinical notes [263]. The challenge of I2B2 2011 was to identify and resolve coreferences expressed in narratives, where different terms refer to the same real-world entity. Coreference annotations meticulously indicate these relationships in order to train models to discern and accurately link references to a common entity. Automatic identification of coreferential links improves understanding of complex clinical relationships. Further development of these technologies should enable the medical community to better understand patient histories, presentations, and treatments through the complete resolution of entities in clinical discussions.
6. *The I2B2 2012 Temporal Relations dataset* was curated with a specific focus on aiding in the identification and understanding of temporally related elements present in clinical notes [264]. The dataset challenge centered on precisely identifying and categorizing the temporal relationships conveyed in narratives, with the goal of understanding when events or actions occurred relative to one another. The notes are meticulously annotated to delineate and classify these time-related connections, supporting the development of models that can accurately discern and organize temporal aspects. Key targets involve discerning whether certain phenomena preceded, followed, or co-occurred based on the documentation. The precise annotations of the dataset and the focus on this significant dimension continue to enable valuable progress on a fundamental but complex clinical modeling task [265].
7. *The I2B2 2014 Deidentification & Heart Disease dataset*: This specialized collection addresses two primary objectives: privacy protection through de-identification and the extraction of heart disease information [266]. For de-identification, the task involves removing personally identifiable information (PII) from medical records while preserving anonymity. The datasets are annotated to highlight sensitive PII to remove. Researchers utilize this resource to develop and assess de-identification systems, ensuring ethical use of data for research by maintaining privacy [267] regarding heart disease extraction, annotations tag mentions, diagnostic details, treatments, and related clinical aspects within notes. Leveraging these guidelines supports building models accurately deriving insights into cardiovascular presentations and management [268].
8. *The I2B2 2018 dataset* is divided into two pivotal tracks: Track 1, focusing on clinical trial cohort selection [269], and Track 2, centered around adverse drug events and medication [270]. Track 1 aims to accurately identify eligible patients for clinical trials using EMR data. The annotations mark records matching various eligibility criteria. Researchers use this resource to develop and evaluate ML models that can efficiently pinpoint suitable trial candidates, expediting the enrollment process. Regarding Track 2, the goal here is to detect and classify mentions of adverse drug reactions and medication information in notes. The annotations highlight such instances in clinical narratives. Researchers apply this dataset to advance NLP techniques specifically for precisely extracting and categorizing adverse events and medication details. This contributes to pharmacovigilance and patient safety.

6.3. Medical QA Datasets

In this section, we present an overview of different medical datasets dedicated to the QA task. Table 5 summarizes some of the key features of several of these datasets which, without claiming to be exhaustive, cover the most frequently studied clinical domains.

Table 5. Summary of medical QA datasets.

Dataset	**Domain**	**Size**	**Data Type**
MEDQUAD	QA	11,230 questions	Clinical questions and answers
MEDMCQA	QA	1310 question–answer pairs	Clinical questions and answers
MEDQA-USMLE	QA	500+ question–answer pairs	Clinical vignettes and questions
MQP	QA	1000 question pairs	Clinical question pairs
CLINIQA-PARA	Paraphrasing	8130 paraphrase pairs	Clinical texts
VQA-RAD	Image QA	6000 question–image pairs	Radiological images and questions
PATHVQA	Pathology image QA	Over 14,000 image–question pairs	Pathology images and questions
PUBMEDQA	Abstract QA	Over 13,000 question–abstract pairs	Biomedical abstracts and questions
VQA-MED-2018, 2019, 2020	Visual QA	Over 8000 image–question–answer triples each year	Medical images, questions, and answers
RADVISDIAL	Conversational agent for radiology	Over 1500 dialogues	Text-based dialogues in radiology domain

1. *MEDQUAD dataset* consolidates authoritative medical knowledge from across the NIH into a substantial question–answer resource. It contains 47,457 pairs compiled from 12 respected NIH websites focused on health topics (e.g., cancer.gov, niddk.nih.gov, GARD, and MEDLINEPlus Health Topics). These sources include the National Cancer Institute, the National Institute of Diabetes and Digestive and Kidney Diseases, and the Genetic and Rare Diseases Information Center databases. Questions fall under 37 predefined categories linked to diseases, drugs, diagnostic tests, and other clinical entities. Examples encompass queries about treatment, diagnosis, and side effects. By synthesizing verified content already maintained by leading NIH organizations, MEDQUAD constructs a rich collection of clinically important questions paired with validated responses. These address a diverse range of issues encountered in biomedical research and practice.
2. *MEDMCQA dataset* contains over 194,000 authentic medical entrance exam multiple-choice questions designed to rigorously evaluate language understanding and reasoning abilities [161]. Sourced from the All India Institute of Medical Sciences (AIIMS) and National Eligibility cum Entrance Test (NEET) PG exams in India, the questions cover 2400 diverse healthcare topics across 21 medical subjects, with an average complexity mirrored in clinical practice. Each sample presents a question stem alongside the correct answer and plausible distractors, requiring models to comprehend language at a depth beyond simple retrieval. Correct selection demonstrates an understanding of semantics, concepts, and logical reasoning across topics. By spanning more than 10 specific reasoning skills, MEDMCQA comprehensively tests a model's language and thought processes in a manner analogous to how human exam takers must demonstrate their clinical knowledge and judgment in order to gain entrance to postgraduate medical programs in India. The large scale, intricacy, and realism of these authentic medical testing scenarios establish MEDMCQA as a leading benchmark for developing assistive technologies with human-level reading comprehension, critical thinking, and decision-making skills. The dataset poses a major challenge that pushes the boundaries of language model abilities.
3. *The MEDQA-USMILE dataset*, introduced by [161], provides 2801 question–answer pairs taken directly from the United States Medical Licensing Examination (USMLE).

As the certification exam required to practice medicine in the United States, the USMLE assesses candidates on an extraordinarily broad and in-depth range of clinical skills. As such, it sets the bar for the level of medical acumen expected of MDs within the U.S. healthcare system. Using authentic samples from this rigorous assessment, MedQA USMILE offers insight into the comprehensive knowledge requirements and analytical abilities assessed. It reflects the extremely high standards of medical training and licensure in the USA. Although it is a preliminary size at present, the use of actual USMLE content gives the dataset validity as a test for medical QA systems looking for capabilities equivalent to those of junior doctors. These capabilities are essential for applications designed to help medical students in the USA. Thanks to its authentic test format and its link to the standards applicable to U.S. doctors, MedQA USMILE provides an important first benchmark, focused solely on the requirements of the U.S. system, and paves the way for more powerful aids for clinical training in the country.

4. *The MQP (Medical Question Pairs) dataset* was compiled manually by MDs to provide examples of medical question pairs, whether similar or not [271]. From a random sample of 1524 authentic patient questions, MDs performed two labeling tasks. First, for each original question, they composed a reworded version retaining an equivalent underlying intention in order to generate a "similar pair". At the same time, using overlapping terminology, they devised a "dissimilar pair" on a related but ultimately inapparent topic. This double-matching process produced, for each initial question, both a semantically coherent reworking and a variant that was superficially related but whose answer was not congruent. Only similar pairs are then used in MQP. By asking doctors to rewrite queries in different styles while retaining consistent meaning, the selected similar question instances give the models the ability to discern medical semantic substance beyond superficial correspondence. Their inclusion enables more rigorous evaluation and enhancement of a system's comprehension capabilities at a deeper level.
5. *The CLINIQPARA dataset* comprises a compendium of patient queries crafted with paraphrasing to progress medical QA from Electronic Health Records [272]. It presents 10,578 uniquely reworded questions sorted into 946 semantically distinct clusters anchored to shared clinical intent. Initially harvested from EMRs, the questions were aggregated to cultivate systems that generalize beyond surface forms to grasp the crux of related inquiries couched in diverse linguistic guises. By merging paraphrased variations tied to an identical underlying semantic substance, the inventory arm models have the prowess to discern core significance notwithstanding syntax. Equipped via CLINIQPARA, AI can learn how a single medical consultation may morph in phrasing while retaining essential import, empowering robust interrogation of noisily documented EMR contents. Its expansive scale and clustering by meaning establish it as paramount for advancing healthcare AI's acuity in apprehending intent beneath variances in how patients may pose an identical essential query.
6. *The VQA-RAD dataset* comprises 3515 manually crafted question–response pairs pertaining to 315 distinctive radiological images [273]. On average, approximately 11 inquiries are posed regarding each individual examination. The questions encapsulate a wide spectrum of clinical constructs and findings that a diagnosing radiologist may want to interrogate or validate within an imaging study. Example themes incorporate anatomical components, anomalies, measurements, diagnoses, and more. By aggregating thousands of question–answer annotations across hundreds of studies, VQA-RAD amasses a sizeable compendium to nurture visual QA (VQA) models. Its scale and granularity lend the resource considerable value to cultivating AI capable of helping radiologists unravel the semantics within medical visuals through an interactive query interface.
7. *PATHVQA* is a seminal resource, representing the first VQA dataset curated specifically for pathology [274]. It contains 32,799 manually formulated questions broadly

covering clinical and morphological issues pertinent to the specialty. These questions correlate to 4998 unique digitized tissue slides, averaging approximately 6–7 inquiries per image. PATHVQA aims to mirror the analytical and diagnostic cognition of pathologists when interpreting whole slide images.

8. *PUBMEDQA* is an essential resource for advancing evidence-based QA in the biomedical field [275]. Its objective is to evaluate the ability of models to extract answers from PUBMED abstracts for clinical questions expressed in a yes/no/medium format. The dataset includes 1000 carefully selected questions, associated with expert annotations, which define the correct answer inferred from the literature for questions such as "Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?". In addition, 61,200 real-life but unlabeled questions enable semi-supervised techniques to develop learning. Over 211,000 artificially generated question–answer summaries are incorporated to deepen the learning pool.
9. *The VQA-MED-2018 dataset* was the first of its kind created specifically for visual QA (VQA) using medical images [276]. It was introduced as part of the ImageCLEF 2018 challenge to allow the testing of VQA models in this new healthcare domain. An automatic rule-driven system first extracted captions from images, simplifying sentences and pinpointing response phrases to seed question formulation and candidate ranking. However, purely algorithmic derivation risks semantic inconsistencies or clinical irrelevance. Two annotators with medical experience thoroughly cross-checked each query and response twice. One round ensured semantic logic, while another judged clinical pertinence to associated visuals. This two-step validation by experienced clinicians guaranteed queries and answers not just made sense logically but genuinely applied to practice. As the pioneering dataset for medical VQA, it thus permits exploring how AI can interpret images and language in healthcare contexts. Thanks to its meticulously verified question–response sets grounded in actual medical images and language, VQA-MED-2018 laid the groundwork for advancing and benchmarking systems to help practitioners through image-based insights.
10. *The VQA-MED-2019 dataset* represented the second iteration of questions and answers related to medical images as part of the ImageCLEF 2019 challenge [277]. Its design aimed to further the investigation started by its predecessor. Drawing inspiration from the question patterns observed in radiology contexts within VQA-RAD, it concentrated on the top four classes of inquiries typically encountered: imaging modalities, anatomical orientations, affected organs, and abnormalities. Some question types, like modality or plane, allowed for categorized responses and could thus be modeled as classification tasks. However, specifying abnormalities required generative models since possibilities were not fixed. This evaluation of different AI problem types paralleled the diverse thinking involved in analyzing scans. Questions also emulated natural consultations. By targeting the most common radiology question categories and building on lessons from previous work, VQA-MED-2019 enhanced the foundation for assessing and advancing systems meant to expedite diagnostic comprehension through combined vision and language processing.
11. *The VQA-MED-2020 dataset* represented the third iteration of this influential initiative, presented as part of ImageCLEF 2020 to advance the answer to medical visual questions [278]. The core corpus consisted of diagnostically relevant images, whose diagnosis was derived directly from visual elements. The questions focused on abnormalities, with 330 frequent conditions selected to ensure minimum prevalence in systematically organized search patterns. This guided combination of visual and linguistic elements established the characteristic VQA assessment framework. In addition, VQA-MED-2020 enriched the landscape by introducing visual question generation, eliciting queries endogenously from radiographic content. Over 1001 associated images provided a basis for 2400 carefully constructed questions, first algorithmically designed and then manually refined. These interdependent tasks aimed to foster more nuanced understanding through contextual synergies between vision and language.

By developing both explanatory and generative faculties, VQA-MED-2020 helped characterize key imaging subtleties while cultivating technologies better prepared to participate in clinical workflows. Its balanced approach to both established and emerging problem spaces iteratively honed assessment of progress towards diagnostically adroit interrogative and descriptive proficiencies—skills paramount to intuitively aiding radiographic decision-making.

12. *The RADVISDIAL dataset* has led the way in the emerging field of visual dialog in medical imaging by introducing the first radiology dialog collection for iterative QA modeling [279]. Drawing on annotated cases from the MIMIC-CXR database accompanied by comprehensive reports, the dataset offered both simulated and authentic consultations. A silver set algorithmically generated multi-round interactions from plain text, while a gold standard captured 100 image-based discussions between expert radiologists under rigorous guidelines. Beyond simple question–answer pairs, this sequential format mimicked the richness of clinical dialogue, posing a more ecologically valid test of the systems' contextual capabilities. The comparisons also revealed the importance of historical context for accuracy—an insight with implications for the development of conversational AI assistants. By establishing visual dialogue as a framework for evaluating the progress of models over a succession of queries, RADVISDIAL laid the foundation for the development of technologies that facilitate nuanced diagnostic reasoning through combined visual and linguistic questions and answers. Its dual-level design also provided a benchmark of synthetic and real-world performance as the field continues to advance. RADVISDIAL therefore opened the way to promising work on sequential and contextual modeling for visual dialogue tasks, such as doctor–patient or radiologist discussions.

7. Foundation Models for Healthcare

Foundation models (FMs) marked a revolution in the world of AI. They are machine learning structures formed from large, often unlabeled datasets [66]. They offer a diverse range of applications, extending from text generation to video editing [280] and including such complex fields as protein folding [281] and robotics [282].

One of the leading models to emerge in this area is ChatGPT from OpenAI. Initially conceived as a linguistic model for predicting the next word in a sequence, it has surprised the scientific community with its multifunctional skills. In particular, its applications in the medical field are remarkable, ranging from assisting doctors with licensing examinations to simplifying radiology reports [132,283].

Rapid progress in this field has attracted the attention of healthcare professionals. However, it is difficult to assess the potential benefits and drawbacks of these advancements in a clinical context. To address this, a recent study of 80 distinct clinical FMs developed from EMR data was undertaken with the aim of understanding and addressing these challenges.

7.1. Concept of FM

Foundation models are based on three key concepts. The first is pre-training, which involves training the model on large datasets. For example, a corpus consists of anonymized patient records, scientific articles, and clinical trial reports. The transformation model is trained on this corpus to acquire a general understanding of the syntax and semantics of a medical language. Its level of understanding is assessed on language comprehension tasks, such as named entity recognition.

Once pre-trained, the model can be fine-tuned to domain- or task-specific data. This is known as the fine-tuning stage. Let us take the example of a model pre-trained on general medical data. It could be fine-tuned based on the analysis of symptoms reported by patients to suggest possible diagnoses. This fine-tuning would be carried out using a dataset corresponding to the symptoms and diagnoses of anonymous patients.

The concept of learning transfer is also essential. It enables the knowledge already acquired by the model to be used to adapt it to related tasks. For example, the patient symptom analysis model could be retrained on data linking symptoms, diagnoses, and effective treatments. The aim would be to automatically associate the right treatments with the proposed diagnoses.

7.2. Clinical FMs

EMR data can be utilized to construct foundation models, which primarily fall into two categories: Clinical Language Models (CLaMs) and Models for EMR (FEMRs).

7.2.1. Clinical Language Models (CLaMs)

CLaMs are a subcategory of large-scale linguistic models specialized in clinical and biomedical texts [284]. They are mainly trained to process and generate this type of medical content, such as extracting drug names from notes or summarizing clinical dialogues [99].

By training them on large quantities of internet text, general-purpose LLMs such as ChatGPT, Bloom, and GPT-4 may offer some utility in clinical settings. However, when it comes to specialized tasks, CLaMs generally demonstrate superior performance.

7.2.2. Foundation Models for EMR (FEMRs)

Another important category of clinical FMs is EMR record integration models, called FMs for Electronic Medical Records (FEMRs) [284]. Unlike the language models previously described, FEMRs can process the structured set of events and clinical data contained in a patient's medical record.

When trained on EMR data, FEMRs generate a vector representation densely encapsulating information relating to the care pathway of each patient. This representation in the form of a high-dimensional vector, often termed "patient embedding", has the property of compactly encapsulating all the information relating to the patient and can then serve as input to various predictive models dedicated to tasks such as anticipating the risk of hospital readmission.

Remarkably, the performances of these secondary predictive models are often superior to those of traditional learning models that do not benefit from the overview provided by the patient embedding. By holistically integrating the multiple dimensions of the medical record, these FEMRs therefore seem to capture the overall health status of patients in a richer way than traditional approaches. Their ability to synthesize the complete clinical history into a dense representation opens promising perspectives for the development of new predictive clinical applications.

8. Discussion

LLMs, such as GPT-4 or LLaMA, have demonstrated significant potential in advancing NLP, and their application in medical diagnosis promises to transform this field.

One of the main strengths of LLMs is their ability to quickly analyze clinical data. A doctor could, for example, provide a model like GPT-4 with a complex set of symptoms and, in return, receive suggestions for possible differential diagnoses. This is even more important when considering that these models can access large, up-to-date medical databases, providing valuable information, especially for rare diseases or complex cases.

The potential of LLMs does not stop at large medical institutions. In remote or underserved areas where access to a specialist may be difficult, an LLM-based tool could serve as a first line of diagnostic advice. Of course, this could never replace real medical advice, but it could effectively guide initial care.

Furthermore, the information burden is a constant challenge in medicine. MDs are inundated with data, and LLMs could play a critical role in filtering and prioritizing this information. At the same time, medical students could benefit from these models as interactive learning tools, allowing them to ask questions and receive detailed answers about various medical conditions.

What is also interesting is the ability of LLMs to process natural language. Patients can simply describe their symptoms, and models like BERT could translate those verbal descriptions into precise medical terms. Going further, by combining LLMs with computer vision technologies, we could envision these models in helping interpret medical images.

However, it is essential to remember that although LLMs offer impressive benefits, they should not be used as the final word in diagnosis. They can serve as valuable assistants, but human clinical judgment is important. Additionally, these models would require extensive validation before widespread adoption to ensure their accuracy in a clinical setting.

Although they offer advantages, the use of LLMs in medical diagnosis also presents serious challenges. First, reliability is a major concern. Although an LLM like GPT-4 can provide insights based on a vast amount of data, it is not immune to errors. For example, a model could misinterpret symptoms described by a patient and suggest an incorrect diagnosis, with potentially serious consequences for the patient's health. Additionally, LLMs' ability to generalize from the data they were trained on can sometimes lead them to make mistakes in situations they did not explicitly "see" during their training.

Then, there is the challenge of overlinking. Although LLMs can be useful as support tools, there is a risk that healthcare professionals will begin to rely too heavily on them to the detriment of their clinical judgment. For example, if an LLM suggests a particular diagnosis based on the information provided, an MD might be tempted to follow that suggestion without questioning it, even if their clinical expertise suggests otherwise.

The issue of data privacy is also crucial. LLMs require huge amounts of data to train. If these data come from actual medical records, this could pose privacy concerns, even if the data are anonymized. It is essential to ensure that sensitive patient information remains protected.

Another challenge is that of transparency and interpretability. Decisions made by LLMs can often appear to be a "black box", meaning that doctors and patients may struggle to understand how a particular diagnosis was suggested. Without a clear explanation, it can be difficult to trust the model's recommendations.

Finally, medical ethics are another area of concern. If an incorrect diagnosis is made based on an LLM's suggestions, who is responsible? The doctor, the hospital, or the creator of the model? These questions require careful consideration before the widespread adoption of LLMs in medical diagnosis.

However, although LLMs have the potential to transform medical diagnosis, it is crucial to approach these challenges with caution and diligence to ensure patient safety and well-being.

8.1. Techniques for Mitigating Ethical Risks

The ethical analysis of LLM models in healthcare is crucial, given their potential impact on various aspects of healthcare. To attenuate the ethical risks associated with these models, several techniques can be employed.

In this field, **differentially private learning** is an important technique for protecting the privacy of the data used in LLM training [285,286]. It ensures that adding or deleting a single data point in the training dataset does not significantly affect the model's output. This reduces the risk of the model revealing sensitive information about the patients from whom the data were collected. A key challenge of differentially private learning is to find a balance between privacy and model accuracy. As the level of confidentiality (i.e., the quantity of noise added) increases, the accuracy of the model tends to decrease.

Improving the **interpretability** of LLMs in healthcare is fundamental to enabling professionals to understand the decision-making process [287,288]. In a sector where decisions have far-reaching consequences, such as healthcare, it is essential to understand the basis of a model's recommendations or predictions. This contributes to the confidence and adoption of these technologies by healthcare professionals. Interpretability techniques aim to explain why and how a model has reached a specific conclusion [289]. This can be done by identifying the characteristics or data that were decisive in the model's decision.

For example, in the case of a diagnosis, the model might indicate which symptoms or test results were most influential in its conclusion. Visual tools such as heat maps make explanations accessible. Improved interpretability contributes to the transparency and reliability of recommendations, encouraging their informed integration without substituting for human judgment [290]. It complements expertise for enriched decision-making. Ultimately, this approach ensures that they not only perform well, but are also understandable and trustworthy by professionals, enabling safer use of AI in healthcare.

LLMs in the healthcare field can involuntarily learn from and propagate biases in training data. To overcome this risk, it is crucial to make regular **audits** to detect and rectify these biases [291]. These audits should include performance analyses on a variety of datasets, and focus on identifying any results that may be unfair or discriminatory.

To ensure that these audits are exhaustive and unbiased, they should be carried out by teams made up of members with diverse profiles and from a variety of disciplines. These teams should include not only medical experts, who have an in-depth understanding of the healthcare field but also researchers specializing in ethics, who can provide a critical perspective on the moral implications of model results. This diversity ensures that different points of view and expertise are taken into account when evaluating models. Such an interdisciplinary approach is essential to ensure that LLMs in healthcare are not only technically competent but also fair and unbiased in their applications. This helps to build confidence in the use of AI in medicine and to ensure that the benefits of these technologies accrue to all individuals in an equitable manner.

Fair learning techniques play a vital role in ensuring that LLM models are fair, unbiased, and representative in healthcare. To achieve this objective, several key strategies can be put in place [292]. First, it is important to perform a preliminary assessment of the data to detect any potential bias and assess the representativity of demographic groups, medical conditions and socio-economic contexts. Next, **datasets should be diversified** by integrating information from diverse sources and populations, ensuring that data are collected from underrepresented or marginalized groups to ensure balanced representation. **Data rebalancing techniques**, such as oversampling minorities or undersampling majority groups, can also be used to balance datasets [293,294]. **Regularization models** can be integrated into the learning process to minimize bias, for example by penalizing predictions that reinforce stereotypes or inequalities [295,296]. **Multi-task learning**, which involves training models on multiple tasks simultaneously, can improve their ability to generalize and reduce single-task-specific biases [297–299]. It is also essential to make rigorous tests on various samples in order to evaluate performance and detect biases in the model's predictions. Finally, it is important to train users on the strengths and limitations of LLMs, emphasizing the importance of their clinical expertise in interpreting the results. Interdisciplinary collaboration, involving ethics experts, health professionals, lawyers, and representatives of diverse communities, can also contribute to the balanced design and evaluation of models in health care.

8.2. Challenges and Future Perspectives

The first major challenge lies in the rigorous management of potential biases contained in the vast medical databases used for training AI algorithms. These data may present biases linked to the gender, age, or ethnic origin of patients; therefore, it is essential to collect information that is diverse and representative of the entire population. Thorough controls must also be put in place to detect any abuse that could unfairly impact certain groups.

The challenge of transparency and interpretability in diagnostic support systems is also crucial. Unlike traditional medical diagnoses requiring a detailed understanding of the reasons leading to conclusions, many current AI algorithms still operate as "black boxes", without the ability to clearly explain their reasoning. However, traceability and complete explainability of the analyses produced are essential in medicine. Significant progress remains to be made in explanatory AI.

The security and safety of AI systems and the highly sensitive medical data they analyze and store are also critical issues. The risks of leaks or malicious attacks aimed at compromising the confidentiality or integrity of this highly private information must be anticipated by deploying solid technical protection measures.

Robust validation of algorithm performance on large and varied populations, as well as their ability to maintain a high level of reliability over time, constitute other important challenges. It will also be necessary to adequately train health professionals in the relevant and safe use of these decision-making tools.

While medical AI shows strong potential, its responsible development requires sustained efforts on these different levels to ensure the trust of patients and practitioners in these technologies.

9. Conclusions

In this paper, we have provided an analysis of the potential impacts of major LLMs on the healthcare sector. We started by introducing LLM architectures like ChatGPT, Bloom, and LLaMA, which are composed of billions of parameters and have demonstrated impressive capabilities in language understanding and generation. These models have the potential to process and generate natural language, which can be used to improve medical diagnosis and patient care as well as accelerate medical research.

We then reviewed recent trends in medical datasets used to train such models, classifying them according to different criteria such as their size, their source, or their subject (patient files, scientific articles, etc.). Using these datasets can help train large language models to better understand the nuances and complexities of clinical language, which can be used to improve decision-making, diagnostic support, or the personalization of treatments.

However, we also highlighted the challenges related to the practical use of large linguistic models in the medical field. One of the main concerns is the ethical implications of using these models, as training them requires large amounts of sensitive medical data. There are concerns about data privacy and security, as well as the potential for these models to perpetuate bias and inaccuracies in medical data, which can have serious consequences for patient care.

Despite these challenges, significant progress is being made in the area of LLMs. With the deployment of LLMs trained on a very large scale with public resources, these models could revolutionize the way we approach patient care and medical research. However, it is important to carefully consider the ethical implications of using these models and work to address these concerns to ensure they are used responsibly and effectively.

Author Contributions: Conceptualization, M.A.A. and K.N.; methodology, K.N. and M.A.A.; validation, K.N. and M.A.A.; formal analysis, K.N. and M.A.A.; writing—original draft preparation, K.N.; writing—review and editing, M.A.A.; funding acquisition, M.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work has been supported in part by the New Brunswick Health Research Foundation (NBHRF) and the New Brunswick Innovation Foundation (NBIF).

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
DL	Deep Learning
LLM	Large Language Model
GPT	Generative Pre-trained Transformer
LLaMA	Large Language Model Meta AI
BLOOM	BigScience Large Open-science Open-access Multilingual Language Model
NLP	Natural Language Processing
RAG	Retrieval Augmented Generation
RL	Reinforcement Learning
RLHF	RL from Human Feedback
DPO	Direct Preference Optimization
PPO	Proximal Policy Optimization
GMAI	Generalist medical AI
RCT	Tandomized controlled trial
LDA	Latent Dirichlet allocation
GloVe	Global Vectors for Word Representation
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Approach
ELMo	Embeddings from Language Models
QA	Question Answering
T5	Text-To-Text Transfer Transformer
RNN	Recurrent Neural Network
CNN	Convolutional Neural Network
BART	Bidirectional and Auto-Regressive Transformer
PaLM	Pathways Language Model
Flan-PaLM	Scaling Instruction-Fine-Tuned Language Models
EMRs	Electronic medical records
MD	Medical doctor
MLTL	Multilevel Transfer Learning Technique
LSTM	Long Short Term Memory
MIMIC	Medical Information Mart for Intensive Car
NLM	National Library of Medicine
NCBI	National Center for Biotechnology Information
MeSH	Medical Subject Headings
BIDMC	Beth Israel Deaconess Medical Center
NUBES	Negation and Uncertainty annotations in Biomedical texts in Spanish
CASI	Clinical Abbreviation Sense Inventory
NLI	Natural Language Inference
SNP	Single Nucleotide Polymorphism
I2B2	Informatics for Integrating Biology and the Bedside
N2C2	National Clinical Language Challenges
PII	Personally Identifiable Information
AIIMS	India Institute of Medical Sciences
NEET	National Eligibility cum Entrance Text
USMLE	United States Medical Licensing Examination
VQA	Visual Questions Answering
MQP	Medical Question Pairs
FM	Foundation model
CLaMs	Clinical Language Models
FEMRs	Foundation Models for EMRs
S2ORC	The Semantic Scholar Open Research Corpus
OPT	Open Pre-trained transformer Language Models
CLEF	Conference and Lab of the Evaluation Forum

References

1. Ye, J.; Chen, X.; Xu, N.; Zu, C.; Shao, Z.; Liu, S.; Cui, Y.; Zhou, Z.; Gong, C.; Shen, Y.; et al. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. *arXiv* **2023**, arXiv:2303.10420.
2. OpenAI.; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
3. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
4. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
5. Iroju, O.G.; Olaleke, J.O. A Systematic Review of Natural Language Processing in Healthcare. *Int. J. Inf. Technol. Comput. Sci.* **2015**, *8*, 44–50. [CrossRef]
6. Hossain, E.; Rana, R.; Higgins, N.; Soar, J.; Barua, P.D.; Pisani, A.R.; Turner, K. Natural language Processing in Electronic Health Records in Relation to Healthcare Decision-making: A Systematic Review. *Comput. Biol. Med.* **2023**, *155*, 106649. [CrossRef] [PubMed]
7. Singleton, J.; Li, C.; Akpunonu, P.D.; Abner, E.L.; Kucharska–Newton, A.M. Using Natural Language Processing to Identify Opioid Use Disorder in Electronic Health Record Data. *Int. J. Med. Inform.* **2023**, *170*, 104963. [CrossRef] [PubMed]
8. Shen, Y.; Heacock, L.; Elias, J.; Hentel, K.D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and other Large Language Models are Double-edged Swords. *Radiology* **2023**, *307*, e230163. [CrossRef] [PubMed]
9. Christensen, L.; Haug, P.; Fiszman, M. MPLUS: A Probabilistic Medical Language Understanding System. In Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain, Phildadelphia, PA, USA, 11 July 2002; pp. 29–36.
10. Wang, D.Q.; Feng, L.Y.; Ye, J.G.; Zou, J.G.; Zheng, Y.F. Accelerating the Integration of ChatGPT and Other Large-scale AI Models Into Biomedical Research and Healthcare. *MedComm-Future Med.* **2023**, *2*, e43. [CrossRef]
11. Schaefer, M.; Reichl, S.; ter Horst, R.; Nicolas, A.M.; Krausgruber, T.; Piras, F.; Stepper, P.; Bock, C.; Samwald, M. Large Language Models are Universal Biomedical Simulators. *bioRxiv* **2023**. [CrossRef]
12. Lederman, A.; Lederman, R.; Verspoor, K. Tasks as needs: Reframing the paradigm of clinical natural language processing research for real-world decision support. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 1810–1817. [CrossRef]
13. Zuheros, C.; Martínez-Cámara, E.; Herrera-Viedma, E.; Herrera, F. Sentiment Analysis based Multi-Person Multi-criteria Decision Making Methodology using Natural Language Processing and Deep Learning for Smarter Decision Aid. Case Study of Restaurant Choice using TripAdvisor Reviews. *Inf. Fusion* **2021**, *68*, 22–36. [CrossRef]
14. Wang, Y.H.; Lin, G.Y. Exploring AI-healthcare Innovation: Natural Language Processing-based Patents Analysis for Technology-driven Roadmapping. *Kybernetes* **2023**, *52*, 1173–1189. [CrossRef]
15. Wang, Y.; Zhao, Y.; Callcut, R.; Petzold, L. Integrating Physiological Time Series and Clinical Notes with Transformer for Early Prediction of Sepsis. *arXiv* **2022**, arXiv:2203.14469.
16. Harrer, S. Attention is Not All You Need: The Complicated Case of Ethically Using Large Language Models in Healthcare and Medicine. *eBioMedicine* **2023**, *90*, 104512. [CrossRef] [PubMed]
17. Hu, M.; Pan, S.; Li, Y.; Yang, X. Advancing Medical Imaging with Language Models: A Journey from n-grams to Chatgpt. *arXiv* **2023**, arXiv:2304.04920.
18. Pivovarov, R.; Elhadad, N. Automated Methods for the Summarization of Electronic Health Records. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 938–947. [CrossRef] [PubMed]
19. Yang, X.; Chen, A.; PourNejatian, N.; Shin, H.C.; Smith, K.E.; Parisien, C.; Compas, C.; Martin, C.; Costa, A.B.; Flores, M.G.; et al. A Large Language Model for Electronic Health Records. *NPJ Digit. Med.* **2022**, *5*, 194. [CrossRef] [PubMed]
20. Tian, S.; Yang, W.; Le Grange, J.M.; Wang, P.; Huang, W.; Ye, Z. Smart Healthcare: Making Medical Care more Intelligent. *Glob. Health J.* **2019**, *3*, 62–65. [CrossRef]
21. Iftikhar, L.; Iftikhar, M.F.; Hanif, M.I. Docgpt: Impact of Chatgpt-3 on Health Services as a Virtual Doctor. *EC Paediatr.* **2023**, *12*, 45–55.
22. KS, N.P.; Sudhanva, S.; Tarun, T.; Yuvraaj, Y.; Vishal, D. Conversational Chatbot Builder–Smarter Virtual Assistance with Domain Specific AI. In Proceedings of the 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 26–28 May 2023; pp. 1–4.
23. Hunter, J.; Freer, Y.; Gatt, A.; Reiter, E.; Sripada, S.; Sykes, C. Automatic Generation of Natural Language Nursing shift Summaries in Neonatal Intensive Care: BT-Nurse. *Artif. Intell. Med.* **2012**, *56*, 157–172. [CrossRef] [PubMed]
24. Abacha, A.B.; Yim, W.W.; Adams, G.; Snider, N.; Yetisgen-Yildiz, M. Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations. In Proceedings of the 5th Clinical Natural Language Processing Workshop, Toronto, ON, Canada, 14 July 2023; pp. 503–513.
25. Thawkar, O.; Shaker, A.; Mullappilly, S.S.; Cholakkal, H.; Anwer, R.M.; Khan, S.; Laaksonen, J.; Khan, F.S. Xraygpt: Chest Radiographs Summarization Using Medical Vision-Language Models. *arXiv* **2023**, arXiv:2306.07971.
26. Phongwattana, T.; Chan, J.H. Automated Extraction and Visualization of Metabolic Networks from Biomedical Literature Using a Large Language Model. *bioRxiv* **2023**. [CrossRef]

27. Tian, S.; Jin, Q.; Yeganova, L.; Lai, P.T.; Zhu, Q.; Chen, X.; Yang, Y.; Chen, Q.; Kim, W.; Comeau, D.C.; et al. Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health. *Briefings Bioinform.* **2024**, *25*, bbad493. [CrossRef] [PubMed]
28. Pal, S.; Bhattacharya, M.; Lee, S.S.; Chakraborty, C. A Domain-Specific Next-Generation Large Language Model (LLM) or ChatGPT is Required for Biomedical Engineering and Research. *Ann. Biomed. Eng.* **2023**, *52*, 451–454. [CrossRef] [PubMed]
29. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. Llava-med: Training a large language-and-vision Assistant for Biomedicine in One Day. *arXiv* **2023**, arXiv:2306.00890.
30. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in Medicine: An Overview of its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. [CrossRef] [PubMed]
31. Moor, M.; Banerjee, O.; Abad, Z.S.H.; Krumholz, H.M.; Leskovec, J.; Topol, E.J.; Rajpurkar, P. Foundation Models for Generalist Medical Artificial Intelligence. *Nature* **2023**, *616*, 259–265. [CrossRef] [PubMed]
32. Li, Y.; Li, Z.; Zhang, K.; Dan, R.; Jiang, S.; Zhang, Y. Chatdoctor: A Medical Chat Model Fine-tuned on llama Model Using Medical Domain Knowledge. *arXiv* **2023**, arXiv:2303.14070.
33. Arif, T.B.; Munaf, U.; Ul-Haque, I. The future of Medical Education and Research: Is ChatGPT a Blessing or Blight in Disguise? *Med. Educ. Online* **2023**, *28*, 2181052. [CrossRef] [PubMed]
34. Bahl, L.; Baker, J.; Cohen, P.; Jelinek, F.; Lewis, B.; Mercer, R. Recognition of Continuously Read Natural Corpus. In Proceedings of the ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing, Tulsa, OK, USA, 10–12 April 1978; Volume 3, pp. 422–424.
35. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
36. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
37. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. In Proceedings of the EMNLP, Doha, Qatar, 25–29 October 2014; Volume 14, pp. 1532–1543.
38. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 2227–2237.
39. Zhuang, L.; Wayne, L.; Ya, S.; Jun, Z. A Robustly Optimized BERT Pre-training Approach with Post-training. In Proceedings of the 20th Chinese National Conference on Computational Linguistics, Huhhot, China, 13–15 August 2021; pp. 1218–1227.
40. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef] [PubMed]
41. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.H.; Jindi, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, 7 June 2019; pp. 72–78.
42. Piñeiro-Martín, A.; García-Mateo, C.; Docío-Fernández, L.; López-Pérez, M.d.C. Ethical Challenges in the Development of Virtual Assistants Powered by Large Language Models. *Electronics* **2023**, *12*, 3170. [CrossRef]
43. Kim, T.; Bae, S.; Kim, H.A.; woo Lee, S.; Hong, H.; Yang, C.; Kim, Y.H. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. *arXiv* **2023**, arXiv:2310.05231.
44. Cazzato, G.; Capuzzolo, M.; Parente, P.; Arezzo, F.; Loizzi, V.; Macorano, E.; Marzullo, A.; Cormio, G.; Ingravallo, G. Chat GPT in Diagnostic Human Pathology: Will It Be Useful to Pathologists? A Preliminary Review with 'Query Session' and Future Perspectives. *AI* **2023**, *4*, 1010–1022. [CrossRef]
45. Wu, P.Y.; Cheng, C.W.; Kaddi, C.D.; Venugopalan, J.; Hoffman, R.; Wang, M.D. Omic and Electronic Health Record Big Data Analytics for Precision Medicine. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 263–273. [PubMed]
46. Gupta, N.S.; Kumar, P. Perspective of Artificial Intelligence in Healthcare Data Management: A Journey Towards Precision Medicine. *Comput. Biol. Med.* **2023**, *162*, 107051. [CrossRef] [PubMed]
47. Liu, S.; Wright, A.P.; Patterson, B.L.; Wanderer, J.P.; Turer, R.W.; Nelson, S.D.; McCoy, A.B.; Sittig, D.F.; Wright, A. Using AI-Generated Suggestions from ChatGPT to Optimize Clinical Decision Support. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1237–1245. [CrossRef] [PubMed]
48. Liu, S.; Wright, A.P.; Patterson, B.L.; Wanderer, J.P.; Turer, R.W.; Nelson, S.D.; McCoy, A.B.; Sittig, D.F.; Wright, A. Assessing the Value of ChatGPT for Clinical Decision Support Optimization. *MedRxiv* **2023**. [CrossRef]
49. Thirunavukarasu, A.J.; Hassan, R.; Mahmood, S.; Sanghera, R.; Barzangi, K.; El Mukashfi, M.; Shah, S. Trialling a Large Language Model (ChatGPT) in General Practice with the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary care. *JMIR Med. Educ.* **2023**, *9*, e46599. [CrossRef] [PubMed]
50. Jo, E.; Epstein, D.A.; Jung, H.; Kim, Y.H. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–16.
51. Lai, T.M.; Zhai, C.; Ji, H. KEBLM: Knowledge-Enhanced Biomedical Language Models. *J. Biomed. Inform.* **2023**, *143*, 104392. [CrossRef] [PubMed]
52. Arsenyan, V.; Bughdaryan, S.; Shaya, F.; Small, K.; Shahnazaryan, D. Large Language Models for Biomedical Knowledge Graph Construction: Information Extraction from EMR Notes. *arXiv* **2023**, arXiv:2301.12473.

53. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
54. Radford, A. Improving Language Understanding by Generative Pre-Training. 2024. Available online: https://openai.com/research/language-unsupervised (accessed on 14 January 2024).
55. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 5485–5551.
56. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* **2019**, *1*, 9.
57. Nassiri, K.; Akhloufi, M. Transformer Models used for Text-based Question Answering Systems. *Appl. Intell.* **2023**, *53*, 10602–10635. [CrossRef]
58. Larochelle, H.; Hinton, G. Learning to Combine Foveal Glimpses with a Third-Order Boltzmann Machine. In Proceedings of the 23rd International Conference on Neural Information Processing Systems—Volume 1, NIPS'10, Vancouver, BC, Canada, 6–9 December 2010; pp. 1243–1251.
59. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, 7–9 May 2015.
60. Luong, T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 1412–1421.
61. Cheng, J.; Dong, L.; Lapata, M. Long Short-Term Memory-Networks for Machine Reading. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 551–561.
62. Parikh, A.; Täckström, O.; Das, D.; Uszkoreit, J. A Decomposable Attention Model for Natural Language Inference. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2249–2255.
63. Paulus, R.; Xiong, C.; Socher, R. A Deep Reinforced Model for Abstractive Summarization. In Proceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, 30 April–3 May 2018.
64. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y. Convolutional Sequence to Sequence Learning. In Proceedings of the Thirty-fourth International Conference on Machine Learning, ICML, Sydney, Australia, 6–11 August 2017.
65. Lin, Z.; Feng, M.; dos Santos, C.N.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A Structured Self-attentive Sentence Embedding. In Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, 24–26 April 2017.
66. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2022**, arXiv:2108.07258.
67. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20), Red Hook, NY, USA, 6–12 December 2020.
68. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
69. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
70. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223.
71. Alajrami, A.; Aletras, N. How does the Pre-training Objective affect what Large Language Models learn about Linguistic Properties? In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Dublin, Ireland, 22–27 May 2022; pp. 131–147.
72. Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; McHardy, R. Challenges and Applications of Large Language Models. *arXiv* **2023**, arXiv:2307.10169.
73. Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. Constitutional AI: Harmlessness from AI Feedback. *arXiv* **2022**, arXiv:2212.08073.
74. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training Language Models to Follow Instructions with Human Feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
75. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent Abilities of Large Language Models. *arXiv* **2022**, arXiv:2206.07682.
76. Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; et al. Solving Quantitative Reasoning Problems with Language Models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 3843–3857.
77. Wang, W.; Zheng, V.W.; Yu, H.; Miao, C. A survey of Zero-shot Learning: Settings, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 1–37. [CrossRef]
78. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought Prompting Elicits Reasoning in Large Language Models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
79. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency Improves Chain of Thought Reasoning in Language Models. *arXiv* **2023**, arXiv:2203.11171.

80. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct preference optimization: Your language model is secretly a reward model. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–12 December 2024; Volume 36.
81. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
82. Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; et al. Zephyr: Direct Distillation of LM Alignment. *arXiv* **2023**, arXiv:2310.16944.
83. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 7–11 November 2021; pp. 3045–3059.
84. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling instruction-finetuned language models. *arXiv* **2022**, arXiv:2210.11416.
85. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
86. Neelakantan, A.; Xu, T.; Puri, R.; Radford, A.; Han, J.M.; Tworek, J.; Yuan, Q.; Tezak, N.; Kim, J.W.; Hallacy, C.; et al. Text and Code Embeddings by Contrastive Pre-Training. *arXiv* **2022**, arXiv:2201.10005.
87. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling Language Modeling with Pathways. *arXiv* **2022**, arXiv:2204.02311.
88. Anil, R.; Dai, A.M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. Palm 2 Technical Report. *arXiv* **2023**, arXiv:2305.10403.
89. Le Scao, T.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; Gallé, M.; et al. Bloom: A 176b-parameter Open-access Multilingual Language Model. *arXiv* **2023**, arXiv:2211.05100.
90. Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; Hashimoto, T.B. Alpaca: A Strong, Replicable Instruction-following Model. *Stanf. Cent. Res. Found. Models.* **2023**, *3*, 7. Available online: https://crfm.stanford.edu/2023/03/13/alpaca.html (accessed on 24 January 2024).
91. Islamovic, A. Stability AI Launches the First of Its StableLM Suite of Language Models-Stability AI. 2023. Available online: https://stability.ai/news/stability-ai-launches-the-first-of-its-stablelm-suite-of-language-models (accessed on 24 January 2024).
92. Conover, M.; Hayes, M.; Mathur, A.; Meng, X.; Xie, J.; Wan, J.; Shah, S.; Ghodsi, A.; Wendell, P.; Zaharia, M.; et al. Free dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. 2023. Available online: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm (accessed on 24 January 2024).
93. Bowman, S.R. Eight Things to Know about Large Language Models. *arXiv* **2023**, arXiv:2304.00612.
94. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.Y. BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining. *Briefings Bioinform.* **2022**, *23*, bbac409. [CrossRef]
95. Bolton, E.; Hall, D.; Yasunaga, M.; Lee, T.; Manning, C.; Liang, P. Stanford CRFM introduces Pubmedgpt 2.7 b. 2022. Available online: https://hai.stanford.edu/news/stanford-crfm-introduces-pubmedgpt-27b (accessed on 24 January 2024).
96. Xiong, H.; Wang, S.; Zhu, Y.; Zhao, Z.; Liu, Y.; Huang, L.; Wang, Q.; Shen, D. DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task. *arXiv* **2023**, arXiv:2304.01097.
97. Chen, Z.; Chen, J.; Zhang, H.; Jiang, F.; Chen, G.; Yu, F.; Wang, T.; Liang, J.; Zhang, C.; Zhang, Z.; et al. LLM Zoo: Democratizing ChatGPT. 2023. Available online: https://github.com/FreedomIntelligence/LLMZoo (accessed on 24 January 2024).
98. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. Towards Expert-level Medical Question Answering with Large Language Models. *arXiv* **2023**, arXiv:2305.09617.
99. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large Language Models Encode Clinical Knowledge. *Nature* **2023**, *620*, 172–180. [CrossRef] [PubMed]
100. Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. Towards Generalist Biomedical AI. *arXiv* **2023**, arXiv:2307.14334.
101. Driess, D.; Xia, F.; Sajjadi, M.S.M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. PaLM-E: An embodied multimodal language model. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
102. Rabe, M.N.; Staats, C. Self-attention Does Not Need $O(n^2)$ Memory. *arXiv* **2022**, arXiv:2112.05682.
103. Korthikanti, V.A.; Casper, J.; Lym, S.; McAfee, L.; Andersch, M.; Shoeybi, M.; Catanzaro, B. Reducing activation recomputation in large transformer models. *arXiv* **2022**, arXiv:2205.05198.
104. Ainslie, J.; Lee-Thorp, J.; de Jong, M.; Zemlyanskiy, Y.; Lebron, F.; Sanghai, S. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 4895–4901.
105. Gema, A.P.; Daines, L.; Minervini, P.; Alex, B. Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain. *arXiv* **2023**, arXiv:2307.03042.
106. Wu, C.; Lin, W.; Zhang, X.; Zhang, Y.; Wang, Y.; Xie, W. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. *arXiv* **2023**, arXiv:2304.14454.
107. Shu, C.; Chen, B.; Liu, F.; Fu, Z.; Shareghi, E.; Collier, N. Visual Med-Alpaca: A Parameter-Efficient Biomedical LLM with Visual Capabilities. 2023. Available online: https://cambridgeltl.github.io/visual-med-alpaca/ (accessed on 24 January 2024).

108. Guevara, M.; Chen, S.; Thomas, S.; Chaunzwa, T.L.; Franco, I.; Kann, B.; Moningi, S.; Qian, J.; Goldstein, M.; Harper, S.; et al. Large Language Models to Identify Social Determinants of Health in Electronic Health Records. *arXiv* **2023**, arXiv:2308.06354.
109. Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; et al. Summary of Chatgpt/gpt-4 Research and Perspective towards the Future of Large Language Models. *Meta-Radiology* **2023**, *1*, 100017. [CrossRef]
110. Wagner, T.; Shweta, F.; Murugadoss, K.; Awasthi, S.; Venkatakrishnan, A.; Bade, S.; Puranik, A.; Kang, M.; Pickering, B.W.; O'Horo, J.C.; et al. Augmented Curation of Clinical Notes from a Massive EHR System Reveals Symptoms of Impending COVID-19 Diagnosis. *eLife* **2020**, *9*, e58227. [CrossRef] [PubMed]
111. Wang, M.; Wang, M.; Yu, F.; Yang, Y.; Walker, J.; Mostafa, J. A systematic review of Automatic Text Summarization for Biomedical Literature and EHRs. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 2287–2297. [CrossRef] [PubMed]
112. Gershanik, E.F.; Lacson, R.; Khorasani, R. Critical Finding Capture in the Impression Section of Radiology Reports. *AMIA Annu. Symp. Proc.* **2011**, *2011*, 465. [PubMed]
113. Choi, E.; Xiao, C.; Stewart, W.; Sun, J. Mime: Multilevel Medical Embedding of Electronic Health Records for Predictive Healthcare. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; Volume 31.
114. Cai, X.; Liu, S.; Han, J.; Yang, L.; Liu, Z.; Liu, T. Chestxraybert: A Pretrained Language Model for Chest Radiology Report Summarization. *IEEE Trans. Multimed.* **2021**, *25*, 845–855. [CrossRef]
115. Xie, Q.; Luo, Z.; Wang, B.; Ananiadou, S. A Survey for Biomedical Text Summarization: From Pre-trained to Large Language Models. *arXiv* **2023**, arXiv:2304.08763.
116. Sharma, A.; Feldman, D.; Jain, A. Team Cadence at MEDIQA-Chat 2023: Generating, Augmenting and Summarizing Clinical Dialogue with Large Language Models. In Proceedings of the 5th Clinical Natural Language Processing Workshop, Toronto, ON, Canada, 14 July 2023; pp. 228–235.
117. Sushil, M.; Kennedy, V.E.; Mandair, D.; Miao, B.Y.; Zack, T.; Butte, A.J. CORAL: Expert-Curated Medical Oncology Reports to Advance Language Model Inference. *arXiv* **2024**, arXiv:2308.03853.
118. Li, H.; Wu, Y.; Schlegel, V.; Batista-Navarro, R.; Nguyen, T.T.; Kashyap, A.R.; Zeng, X.; Beck, D.; Winkler, S.; Nenadic, G. PULSAR: Pre-training with Extracted Healthcare Terms for Summarizing Patients' Problems and Data Augmentation with Black-box Large Language Models. *arXiv* **2023**, arXiv:2306.02754.
119. Park, G.; Yoon, B.J.; Luo, X.; Lpez-Marrero, V.; Johnstone, P.; Yoo, S.; Alexander, F. Automated Extraction of Molecular Interactions and Pathway Knowledge using Large Language Model, Galactica: Opportunities and Challenges. In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, ON, Canada, 13 July 2023; pp. 255–264.
120. Kartchner, D.; Ramalingam, S.; Al-Hussaini, I.; Kronick, O.; Mitchell, C. Zero-Shot Information Extraction for Clinical Meta-Analysis using Large Language Models. In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, ON, Canada, 13 July 2023; pp. 396–405.
121. Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; Sontag, D. Large Language Models are Few-shot Clinical Information Extractors. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 1998–2022.
122. Wu, J.; Shi, D.; Hasan, A.; Wu, H. KnowLab at RadSum23: Comparing Pre-trained Language Models in Radiology Report Summarization. In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, ON, Canada, 13 July 2023; pp. 535–540.
123. Yan, A.; McAuley, J.; Lu, X.; Du, J.; Chang, E.Y.; Gentili, A.; Hsu, C.N. RadBERT: Adapting Transformer-based Language Models to Radiology. *Radiol. Artif. Intell.* **2022**, *4*, e210258. [CrossRef] [PubMed]
124. Dash, D.; Thapa, R.; Banda, J.M.; Swaminathan, A.; Cheatham, M.; Kashyap, M.; Kotecha, N.; Chen, J.H.; Gombar, S.; Downing, L. Evaluation of GPT-3.5 and GPT-4 for Supporting Real-World Information Needs in Healthcare Delivery. *arXiv* **2023**, arXiv:2304.13714.
125. Li, H.; Gerkin, R.C.; Bakke, A.; Norel, R.; Cecchi, G.; Laudamiel, C.; Niv, M.Y.; Ohla, K.; Hayes, J.E.; Parma, V.; et al. Text-based Predictions of COVID-19 Diagnosis from Self-reported Chemosensory Descriptions. *Commun. Med.* **2023**, *3*, 104. [CrossRef] [PubMed]
126. Liu, M.; Zhang, D.; Tan, W.; Zhang, H. DeakinNLP at ProbSum 2023: Clinical Progress Note Summarization with Rules and Language ModelsClinical Progress Note Summarization with Rules and Languague Models. In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, ON, Canada, 13 July 2023; pp. 491–496.
127. Macharla, S.; Madamanchi, A.; Kancharla, N. nav-nlp at RadSum23: Abstractive Summarization of Radiology Reports using BART Finetuning. In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, ON, Canada, 13 July 2023; pp. 541–544.
128. Koga, S.; Martin, N.B.; Dickson, D.W. Evaluating the Performance of Large Language Models: ChatGPT and Google Bard in Generating Differential Diagnoses in Clinicopathological Conferences of Neurodegenerative Disorders. *Brain Pathol.* 2023, e13207, *early view*.
129. Balas, M.; Ing, E.B. Conversational AI Models for Ophthalmic Diagnosis: Comparison of Chatbot and the Isabel pro Differential Diagnosis Generator. *JFO Open Ophthalmol.* **2023**, *1*, 100005. [CrossRef]
130. Huang, H.; Zheng, O.; Wang, D.; Yin, J.; Wang, Z.; Ding, S.; Yin, H.; Xu, C.; Yang, R.; Zheng, Q.; et al. ChatGPT for Shaping the Future of Dentistry: The Potential of Multi-modal Large Language Model. *Int. J. Oral Sci.* **2023**, *15*, 29. [CrossRef] [PubMed]
131. Zhong, Y.; Chen, Y.J.; Zhou, Y.; Yin, J.J.; Gao, Y.J. The Artificial Intelligence Large Language Models and Neuropsychiatry Practice and Research Ethic. *Asian J. Psychiatry* **2023**, *84*, 103577. [CrossRef] [PubMed]

132. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education using Large Language Models. *PLoS Digit. Health* **2023**, *2*, e0000198. [CrossRef] [PubMed]
133. Eggmann, F.; Weiger, R.; Zitzmann, N.U.; Blatz, M.B. Implications of Large Language Models such as ChatGPT for Dental Medicine. *J. Esthet. Restor. Dent.* **2023**, *35*, 1098–1102. [CrossRef] [PubMed]
134. Lehman, E.; Johnson, A. Clinical-t5: Large Language Models Built Using Mimic Clinical Text. 2023. Available online: https://www.physionet.org/content/clinical-t5/1.0.0/ (accessed on 24 January 2024).
135. Ma, C.; Wu, Z.; Wang, J.; Xu, S.; Wei, Y.; Liu, Z.; Jiang, X.; Guo, L.; Cai, X.; Zhang, S.; et al. ImpressionGPT: An Iterative Optimizing Framework for Radiology Report Summarization with ChatGPT. *arXiv* **2023**, arXiv:2304.08448.
136. Liu, Z.; Zhong, A.; Li, Y.; Yang, L.; Ju, C.; Wu, Z.; Ma, C.; Shu, P.; Chen, C.; Kim, S.; et al. Radiology-GPT: A Large Language Model for Radiology. *arXiv* **2023**, arXiv:2306.08666.
137. Li, C.; Zhang, Y.; Weng, Y.; Wang, B.; Li, Z. Natural Language Processing Applications for Computer-Aided Diagnosis in Oncology. *Diagnostics* **2023**, *13*, 286. [CrossRef] [PubMed]
138. Joseph, S.A.; Chen, L.; Trienes, J.; Göke, H.L.; Coers, M.; Xu, W.; Wallace, B.C.; Li, J.J. FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence. *arXiv* **2024**, arXiv:2402.11456.
139. Van Veen, D.; Van Uden, C.; Attias, M.; Pareek, A.; Bluethgen, C.; Polacin, M.; Chiu, W.; Delbrouck, J.B.; Zambrano Chaves, J.; Langlotz, C.; et al. RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models. In Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, Toronto, ON, Canada, 13 July 2023; pp. 449–460.
140. Den Hamer, D.M.; Schoor, P.; Polak, T.B.; Kapitan, D. Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models. *arXiv* **2023**, arXiv:2304.07396.
141. Cascella, M.; Montomoli, J.; Bellini, V.; Bignami, E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J. Med. Syst.* **2023**, *47*, 33. [CrossRef] [PubMed]
142. Tang, L.; Peng, Y.; Wang, Y.; Ding, Y.; Durrett, G.; Rousseau, J. Less Likely Brainstorming: Using Language Models to Generate Alternative Hypotheses. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 12532–12555.
143. Chen, S.; Wu, M.; Zhu, K.Q.; Lan, K.; Zhang, Z.; Cui, L. LLM-empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation. *arXiv* **2023**, arXiv:2305.13614.
144. Kleesiek, J.; Wu, Y.; Stiglic, G.; Egger, J.; Bian, J. An Opinion on ChatGPT in Health Care—Written by Humans Only. *J. Nucl. Med.* **2023**, *64*, 701–703. [CrossRef] [PubMed]
145. Jackson, R.G.; Patel, R.; Jayatilleke, N.; Kolliakou, A.; Ball, M.; Gorrell, G.; Roberts, A.; Dobson, R.J.; Stewart, R. Natural Language Processing to Extract Symptoms of Severe Mental Illness from Clinical Text: The Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* **2017**, *7*, e012012. [CrossRef] [PubMed]
146. Liang, S.; Hartmann, M.; Sonntag, D. Cross-domain German Medical Named Entity Recognition using a Pre-Trained Language Model and Unified Medical Semantic Types. In Proceedings of the 5th Clinical Natural Language Processing Workshop, Toronto, ON, Canada, 14 July 2023; pp. 259–271.
147. Harskamp, R.E.; De Clercq, L. Performance of ChatGPT as an AI-assisted Decision Support Tool in Medicine: A Proof-of-concept Study for Interpreting Symptoms and Management of Common Cardiac Conditions (AMSTELHEART-2). *Acta Cardiol.* **2024**, 1–9. [CrossRef] [PubMed]
148. Rao, A.; Kim, J.; Kamineni, M.; Pang, M.; Lie, W.; Dreyer, K.J.; Succi, M.D. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J. Am. Coll. Radiol.* **2023**, *20*, 990–997. [CrossRef] [PubMed]
149. Lyu, C.; Wu, M.; Wang, L.; Huang, X.; Liu, B.; Du, Z.; Shi, S.; Tu, Z. Macaw-LLM: Multi-Modal Language Modeling with Image, Audio, Video, and Text Integration. *arXiv* **2023**, arXiv:2306.09093.
150. Drozdov, I.; Forbes, D.; Szubert, B.; Hall, M.; Carlin, C.; Lowe, D.J. Supervised and Unsupervised Language Modelling in Chest X-ray Radiological Reports. *PLoS ONE* **2020**, *15*, e0229963. [CrossRef] [PubMed]
151. Nath, C.; Albaghdadi, M.S.; Jonnalagadda, S.R. A Natural Language Processing Tool for Large-scale Data Extraction from Echocardiography Reports. *PLoS ONE* **2016**, *11*, e0153749. [CrossRef] [PubMed]
152. Naseem, U.; Bandi, A.; Raza, S.; Rashid, J.; Chakravarthi, B.R. Incorporating Medical Knowledge to Transformer-based Language Models for Medical Dialogue Generation. In Proceedings of the 21st Workshop on Biomedical Language Processing, Dublin, Ireland, 26 May 2022; pp. 110–115.
153. Zhou, H.Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; Li, W. A Transformer-based Representation-Learning Model with Unified Processing of Multimodal Input for Clinical Diagnostics. *Nat. Biomed. Eng.* **2023**, *7*, 743–755. [CrossRef] [PubMed]
154. Biswas, S. ChatGPT and the future of medical writing. *Radiology* **2023**, *307*, e223312. [CrossRef] [PubMed]
155. Shortliffe, E.H. Computer Programs to Support Clinical decision making. *JAMA* **1987**, *258*, 61–66. [CrossRef] [PubMed]
156. Szolovits, P.; Pauker, S.G. Categorical and probabilistic reasoning in medicine revisited. *Artif. Intell.* **1993**, *59*, 167–180. [CrossRef]
157. Yasunaga, M.; Leskovec, J.; Liang, P. LinkBERT: Pretraining Language Models with Document Links. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 8003–8016.

158. Yasunaga, M.; Bosselut, A.; Ren, H.; Zhang, X.; Manning, C.D.; Liang, P.S.; Leskovec, J. Deep Bidirectional Language-Knowledge Graph Pretraining. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 37309–37323.
159. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthc. (HEALTH)* **2021**, *3*, 1–23. [CrossRef]
160. Jin, D.; Pan, E.; Oufattole, N.; Weng, W.H.; Fang, H.; Szolovits, P. What Disease does this Patient have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *Appl. Sci.* **2021**, *11*, 6421. [CrossRef]
161. Pal, A.; Umapathi, L.K.; Sankarasubbu, M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In Proceedings of the Conference on Health, Inference, and Learning, Virtual Event, 7–8 April 2022; Volume 174, pp. 248–260.
162. Zhu, X.; Chen, Y.; Gu, Y.; Xiao, Z. SentiMedQAer: A Transfer Learning-Based Sentiment-Aware Model for Biomedical Question Answering. *Front. Neurorobot.* **2022**, *16*, 773329. [CrossRef] [PubMed]
163. Zhou, S.; Zhang, Y. Datlmedqa: A Data Augmentation and Transfer Learning based Solution for Medical Question Answering. *Appl. Sci.* **2021**, *11*, 11251. [CrossRef]
164. Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; et al. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. In Proceedings of the BigScience Episode #5—Workshop on Challenges & Perspectives in Creating Large Language Models, Virtual, 27 May 2022; pp. 95–136.
165. Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X.V.; et al. Opt: Open Pre-trained Transformer Language Models. *arXiv* **2022**, arXiv:2205.01068.
166. Rosol, M.; Gasior, J.S.; Laba, J.; Korzeniewski, K.; Młyńczak, M. Evaluation of the Performance of GPT-3.5 and GPT-4 on the Medical Final Examination. *Sci. Rep.* **2023**, *13*, 20512. [CrossRef] [PubMed]
167. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of gpt-4 on Medical Challenge Problems. *arXiv* **2023**, arXiv:2303.13375.
168. Nashwan, A.J.; Abujaber, A.A.; Choudry, H. Embracing the Future of Physician-patient Communication: GPT-4 in Gastroenterology. *Gastroenterol. Endosc.* **2023**, *1*, 132–135. [CrossRef]
169. Meskó, B.; Topol, E.J. The Imperative for Regulatory Oversight of Large Language Models (or Generative AI) in Healthcare. *NPJ Digit. Med.* **2023**, *6*, 120. [CrossRef] [PubMed]
170. Shin, D. The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI. *Int. J. Hum.-Comput. Stud.* **2021**, *146*, 102551. [CrossRef]
171. Khowaja, S.A.; Khuwaja, P.; Dev, K. ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) Evaluation: A Review. *arXiv* **2023**, arXiv:2305.03123.
172. Ferrara, E. Should Chatgpt be Biased? Challenges and Risks of Bias in Large Language Models. *arXiv* **2023**, arXiv:2304.03738.
173. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
174. Zakka, C.; Shad, R.; Chaurasia, A.; Dalal, A.R.; Kim, J.L.; Moor, M.; Fong, R.; Phillips, C.; Alexander, K.; Ashley, E.; et al. Almanac—Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI* **2024**, *1*, AIoa2300068. [CrossRef]
175. Liu, N.; Zhang, T.; Liang, P. Evaluating Verifiability in Generative Search Engines. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 7001–7025.
176. Jin, Q.; Leaman, R.; Lu, Z. Retrieve, Summarize, and Verify: How will ChatGPT impact information seeking from the medical literature? *J. Am. Soc. Nephrol.* **2023**, *34*, 10–1681. [CrossRef] [PubMed]
177. Parisi, A.; Zhao, Y.; Fiedel, N. TALM: Tool Augmented Language Models. *arXiv* **2022**, arXiv:2205.12255.
178. Jin, Q.; Yang, Y.; Chen, Q.; Lu, Z. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics* **2024**, *40*, btae075. [CrossRef] [PubMed]
179. Qin, Y.; Hu, S.; Lin, Y.; Chen, W.; Ding, N.; Cui, G.; Zeng, Z.; Huang, Y.; Xiao, C.; Han, C.; et al. Tool Learning with Foundation Models. *arXiv* **2023**, arXiv:2304.08354.
180. Gao, L.; Madaan, A.; Zhou, S.; Alon, U.; Liu, P.; Yang, Y.; Callan, J.; Neubig, G. PAL: Program-aided language models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
181. Farhadi, A.; Hejrati, M.; Sadeghi, M.A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; Forsyth, D. Every Picture Tells a Story: Generating Sentences from Images. In Proceedings of the Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Proceedings, Part IV 11; pp. 15–29.
182. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. Babytalk: Understanding and Generating Simple Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2891–2903. [CrossRef]
183. Li, S.; Kulkarni, G.; Berg, T.; Berg, A.; Choi, Y. Composing Simple Image Descriptions using Web-scale N-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Portland, OR, USA, 23–24 June 2011; pp. 220–228.
184. Yao, B.Z.; Yang, X.; Lin, L.; Lee, M.W.; Zhu, S.C. I2t: Image Parsing to Text Description. *Proc. IEEE* **2010**, *98*, 1485–1508. [CrossRef]
185. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 652–663. [CrossRef] [PubMed]
186. Johnson, J.; Karpathy, A.; Fei-Fei, L. Densecap: Fully Convolutional Localization Networks for Dense Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4565–4574.

187. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Hierarchy Parsing for Image Captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2621–2629.
188. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-Encoding Scene Graphs for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10685–10694.
189. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring Visual Relationship for Image Captioning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 684–699.
190. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 375–383.
191. Li, X.; Jiang, S. Know More Say Less: Image Captioning Based on Scene graphs. *IEEE Trans. Multimed.* **2019**, *21*, 2117–2130. [CrossRef]
192. Chunseong Park, C.; Kim, B.; Kim, G. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 895–903.
193. Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W.W.; Salakhutdinov, R.R. Review Networks for Caption Generation. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
194. Wang, Y.; Lin, Z.; Shen, X.; Cohen, S.; Cottrell, G.W. Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7272–7281.
195. Wang, Q.; Chan, A.B. CNN+ CNN: Convolutional Decoders for Image Captioning. *arXiv* **2018**, arXiv:1805.09019.
196. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10578–10587.
197. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
198. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked Cross Attention for Image-Text Matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 201–216.
199. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled Transformer for Image Captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8928–8937.
200. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.
201. Alsharid, M.; Sharma, H.; Drukker, L.; Chatelain, P.; Papageorghiou, A.T.; Noble, J.A. Captioning Ultrasound Images Automatically. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019; Proceedings, Part IV 22; Springer: Berlin/Heidelberg, Germany, 2019; pp. 338–346.
202. Alsharid, M.; El-Bouri, R.; Sharma, H.; Drukker, L.; Papageorghiou, A.T.; Noble, J.A. A Curriculum Learning based Approach to Captioning Ultrasound Images. In Proceedings of the Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis: First International Workshop, ASMUS 2020, and 5th International Workshop, PIPPI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, 4–8 October 2020; Proceedings 1; Springer: Berlin/Heidelberg, Germany, 2020; pp. 75–84.
203. Alsharid, M.; El-Bouri, R.; Sharma, H.; Drukker, L.; Papageorghiou, A.T.; Noble, J.A. A Course-focused Dual Curriculum for Image Captioning. In Proceedings of the IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; pp. 716–720.
204. Aswiga, R.; Shanthi, A. A Multilevel Transfer Learning Technique and LSTM Framework for Generating Medical Captions for Limited CT and DBT Images. *J. Digit. Imaging* **2022**, *35*, 564–580. [CrossRef] [PubMed]
205. Selivanov, A.; Rogov, O.Y.; Chesakov, D.; Shelmanov, A.; Fedulova, I.; Dylov, D.V. Medical Image Captioning via Generative Pretrained Transformers. *Sci. Rep.* **2023**, *13*, 4171. [CrossRef] [PubMed]
206. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Volume 37, pp. 2048–2057.
207. Alsharid, M.; Cai, Y.; Sharma, H.; Drukker, L.; Papageorghiou, A.T.; Noble, J.A. Gaze-Assisted Automatic Captioning of Fetal Ultrasound Videos using Three-way Multi-modal Deep Neural Networks. *Med. Image Anal.* **2022**, *82*, 102630. [CrossRef] [PubMed]
208. Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; Qiao, Y. VideoChat: Chat-Centric Video Understanding. *arXiv* **2024**, arXiv:2305.06355.
209. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.w.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; Mark, R.G. MIMIC-III, a Freely Accessible Critical Care Database. *Sci. Data* **2016**, *3*, 160035. [CrossRef] [PubMed]
210. Zhu, Y.; Zhang, J.; Wang, G.; Yao, R.; Ren, C.; Chen, G.; Jin, X.; Guo, J.; Liu, S.; Zheng, H.; et al. Machine Learning Prediction Models for Mechanically Ventilated Patients: Analyses of the MIMIC-III Database. *Front. Med.* **2021**, *8*, 662340. [CrossRef] [PubMed]
211. Khope, S.R.; Elias, S. Simplified & Novel Predictive Model using Feature Engineering over MIMIC-III Dataset. *Procedia Comput. Sci.* **2023**, *218*, 1968–1976.

212. Huang, H.; Liu, Y.; Wu, M.; Gao, Y.; Yu, X. Development and Validation of a Risk Stratification Model for Predicting the Mortality of Acute Kidney Injury in Critical Care Patients. *Ann. Transl. Med.* **2021**, *9*, 323. [CrossRef]
213. Li, D.; Gao, J.; Hong, N.; Wang, H.; Su, L.; Liu, C.; He, J.; Jiang, H.; Wang, Q.; Long, Y.; et al. A Clinical Prediction Model to Predict Heparin Treatment Outcomes and Provide Dosage Recommendations: Development and Validation Study. *J. Med. Internet Res.* **2021**, *23*, e27118. [CrossRef] [PubMed]
214. Khope, S.; Elias, S. Strategies of Predictive Schemes and Clinical Diagnosis for Prognosis Using MIMIC-III: A Systematic Review. *Healthcare* **2023**, *11*, 710. [CrossRef] [PubMed]
215. Wang, R.; Cai, L.; Liu, Y.; Zhang, J.; Ou, X.; Xu, J. Machine Learning Algorithms for Prediction of Ventilator Associated Pneumonia in Traumatic Brain Injury Patients from the MIMIC-III Database. *Heart Lung* **2023**, *62*, 225–232. [CrossRef] [PubMed]
216. Geri, G.; Ferrer, L.; Tran, N.; Celi, L.A.; Jamme, M.; Lee, J.; Vieillard-Baron, A. Cardio-Pulmonary-Renal Interactions in ICU Patients. Role of Mechanical Ventilation, Venous Congestion and Perfusion Deficit on Worsening of Renal Function: Insights from the MIMIC-III Database. *J. Crit. Care* **2021**, *64*, 100–107. [CrossRef] [PubMed]
217. Kurniati, A.P.; Hall, G.; Hogg, D.; Johnson, O. Process Mining in Oncology Using the MIMIC-III Dataset. *J. Phys. Conf. Ser.* **2018**, *971*, 012008. [CrossRef]
218. McWilliams, C.J.; Lawson, D.J.; Santos-Rodriguez, R.; Gilchrist, I.D.; Champneys, A.; Gould, T.H.; Thomas, M.J.; Bourdeaux, C.P. Towards a Decision Support Tool for Intensive Care Discharge: Machine Learning Algorithm Development Using Electronic Healthcare Data from MIMIC-III and Bristol, UK. *BMJ Open* **2019**, *9*, e025925. [CrossRef] [PubMed]
219. Ding, E.Y.; Albuquerque, D.; Winter, M.; Binici, S.; Piche, J.; Bashar, S.K.; Chon, K.; Walkey, A.J.; McManus, D.D. Novel Method of Atrial Fibrillation Case Identification and Burden Estimation Using the MIMIC-III Electronic Health Data Set. *J. Intensive Care Med.* **2019**, *34*, 851–857. [CrossRef]
220. Aldughayfiq, B.; Ashfaq, F.; Jhanjhi, N.Z.; Humayun, M. Capturing Semantic Relationships in Electronic Health Records Using Knowledge Graphs: An Implementation Using MIMIC III Dataset and GraphDB. *Healthcare* **2023**, *11*, 1762. [CrossRef] [PubMed]
221. Zhu, J.L.; Hong, L.; Yuan, S.Q.; Xu, X.M.; Wei, J.R.; Yin, H.Y. Association Between Glucocorticoid Use and All-cause Mortality in Critically Ill Patients with Heart Failure: A Cohort Study Based on the MIMIC-III Database. *Front. Pharmacol.* **2023**, *14*, 1118551. [CrossRef] [PubMed]
222. Johnson, A.E.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.y.; Mark, R.G.; Horng, S. MIMIC-CXR, a De-identified Publicly Available Database of Chest Radiographs with Free-Text Reports. *Sci. Data* **2019**, *6*, 317. [CrossRef] [PubMed]
223. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. Natural Questions: A Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 453–466. [CrossRef]
224. Irmici, G.; Cè, M.; Caloro, E.; Khenkina, N.; Della Pepa, G.; Ascenti, V.; Martinenghi, C.; Papa, S.; Oliva, G.; Cellina, M. Chest X-ray in Emergency Radiology: What Artificial Intelligence Applications Are Available? *Diagnostics* **2023**, *13*, 216. [CrossRef] [PubMed]
225. van Sonsbeek, T.; Worring, M. Towards Automated Diagnosis with Attentive Multi-modal Learning Using Electronic Health Records and Chest X-rays. In Proceedings of the Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures: 10th International Workshop, ML-CDS 2020, and 9th International Workshop, CLIP 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, 4–8 October 2020; Proceedings 9; pp. 106–114.
226. Lin, M.; Hou, B.; Mishra, S.; Yao, T.; Huo, Y.; Yang, Q.; Wang, F.; Shih, G.; Peng, Y. Enhancing Thoracic Disease Detection Using Chest X-rays from PubMed Central Open Access. *Comput. Biol. Med.* **2023**, *159*, 106962. [CrossRef] [PubMed]
227. Glocker, B.; Jones, C.; Bernhardt, M.; Winzeck, S. Algorithmic Encoding of Protected Characteristics in Chest X-ray Disease Detection Models. *eBioMedicine* **2023**, *89*, 104467. [CrossRef] [PubMed]
228. Park, H.; Kim, K.; Park, S.; Choi, J. Medical Image Captioning Model to Convey More Details: Methodological Comparison of Feature Difference Generation. *IEEE Access* **2021**, *9*, 150560–150568. [CrossRef]
229. Légaré, F.; Turcotte, S.; Stacey, D.; Ratté, S.; Kryworuchko, J.; Graham, I.D. Patients' Perceptions of Sharing in Decisions: A Systematic Review of Interventions to Enhance Shared Decision Making in Routine Clinical Practice. *Patient-Patient Outcomes Res.* **2012**, *5*, 1–19. [CrossRef]
230. Barradell, A.C.; Gerlis, C.; Houchen-Wolloff, L.; Bekker, H.L.; Robertson, N.; Singh, S.J. Systematic Review of Shared Decision-making Interventions for People Living with Chronic Respiratory Diseases. *BMJ Open* **2023**, *13*, e069461. [CrossRef]
231. Alpi, K.M.; Martin, C.L.; Plasek, J.M.; Sittig, S.; Smith, C.A.; Weinfurter, E.V.; Wells, J.K.; Wong, R.; Austin, R.R. Characterizing Terminology Applied by Authors and Database Producers to Informatics Literature on Consumer Engagement with Wearable Devices. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1284–1292. [CrossRef]
232. Yuan, P.; Qi, G.; Hu, X.; Qi, M.; Zhou, Y.; Shi, X. Characteristics, Likelihood and Challenges of Road Traffic Injuries in China before COVID-19 and in the Postpandemic Era. *Humanit. Soc. Sci. Commun.* **2023**, *10*, 2. [CrossRef]
233. Rohanian, O.; Nouriborji, M.; Kouchaki, S.; Clifton, D.A. On the Effectiveness of Compact Biomedical Transformers. *Bioinformatics* **2023**, *39*, btad103. [CrossRef] [PubMed]
234. Jimeno Yepes, A.J.; Verspoor, K. Classifying Literature Mentions of Biological Pathogens as Experimentally Studied Using Natural Language Processing. *J. Biomed. Semant.* **2023**, *14*, 1. [CrossRef] [PubMed]

235. Gupta, V.; Dixit, A.; Sethi, S. An Improved Sentence Embeddings based Information Retrieval Technique using Query Reformulation. In Proceedings of the 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, 5–6 May 2023; pp. 299–304.
236. Bascur, J.P.; Verberne, S.; van Eck, N.J.; Waltman, L. Academic Information Retrieval using Citation Clusters: In-Depth Evaluation based on Systematic Reviews. *Scientometrics* **2023**, *128*, 2895–2921. [CrossRef]
237. Bramley, R.; Howe, S.; Marmanis, H. Notes on the Data Quality of Bibliographic Records from the MEDLINE Database. *Database* **2023**, *2023*, baad070. [CrossRef] [PubMed]
238. Chen, M.L.; Rotemberg, V.; Lester, J.C.; Novoa, R.A.; Chiou, A.S.; Daneshjou, R. Evaluation of Diagnosis Diversity in Artificial Intelligence Datasets: A Scoping Review. *Br. J. Dermatol.* **2023**, *188*, 292–294. [CrossRef]
239. Lunge, S.B.; Shetty, N.S.; Sardesai, V.R.; Karagaiah, P.; Yamauchi, P.S.; Weinberg, J.M.; Kircik, L.; Giulini, M.; Goldust, M. Therapeutic Application of Machine Learning in Psoriasis: A Prisma Systematic Review. *J. Cosmet. Dermatol.* **2023**, *22*, 378–382. [CrossRef]
240. Ernst, P.; Siu, A.; Milchevski, D.; Hoffart, J.; Weikum, G. DeepLife: An Entity-aware Search, Analytics and Exploration Platform for Health and Life Sciences. In Proceedings of the ACL-2016 System Demonstrations, Berlin, Germany, 7–12 August 2016; pp. 19–24.
241. Chen, Y.; Elenee Argentinis, J.; Weber, G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clin. Ther.* **2016**, *38*, 688–701. [CrossRef] [PubMed]
242. Haynes, R.; McKibbon, K.; Walker, C.; Mousseau, J.; Baker, L.; Fitzgerald, D.; Guyatt, G.; Norman, G. Computer Searching of the Medical Literature. An Evaluation of MEDLINE Searching Systems. *Ann. Intern. Med.* **1985**, *103*, 812–816. [CrossRef] [PubMed]
243. Stevenson, M.; Guo, Y.; Alamri, A.; Gaizauskas, R. Disambiguation of Biomedical Abbreviations. In Proceedings of the BioNLP 2009 Workshop, Boulder, CO, USA, 4–5 June 2009; pp. 71–79.
244. Lima Lopez, S.; Perez, N.; Cuadros, M.; Rigau, G. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 5772–5781.
245. Doğan, R.I.; Leaman, R.; Lu, Z. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *J. Biomed. Inform.* **2014**, *47*, 1–10. [CrossRef] [PubMed]
246. Subramanian, S.; Wang, L.L.; Bogin, B.; Mehta, S.; van Zuylen, M.; Parasa, S.; Singh, S.; Gardner, M.; Hajishirzi, H. MedICaT: A Dataset of Medical Images, Captions, and Textual References. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 16–20 November 2020; pp. 2112–2120.
247. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.; Thoma, G.R.; McDonald, C.J. Preparing a Collection of Radiology Examinations for Distribution and Retrieval. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 304–310. [CrossRef] [PubMed]
248. Schopf, T.; Braun, D.; Matthes, F. Evaluating Unsupervised Text Classification: Zero-Shot and Similarity-Based Approaches. In Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval, New York, NY, USA, 11–12 July 2023; pp. 6–15.
249. Uzuner, O.; Luo, Y.; Szolovits, P. Evaluating the State-of-the-Art in Automatic De-identification. *J. Am. Med. Inform. Assoc.* **2007**, *14*, 550–563. [CrossRef] [PubMed]
250. Uzuner, Z.; Goldstein, I.; Luo, Y.; Kohane, I. Identifying Patient Smoking Status from Medical Discharge Records. *J. Am. Med. Inform. Assoc.* **2008**, *15*, 14–24. [CrossRef] [PubMed]
251. Uzuner, O. Recognizing Obesity and Comorbidities in Sparse Data. *J. Am. Med. Inform. Assoc.* **2009**, *16*, 561–570. [CrossRef] [PubMed]
252. Liu, S.; Luo, Y.; Stone, D.; Zong, N.; Wen, A.; Yu, Y.; Rasmussen, L.V.; Wang, F.; Pathak, J.; Liu, H.; et al. Integration of NLP2FHIR Representation with Deep Learning Models for EHR Phenotyping: A Pilot Study on Obesity Datasets. *AMIA Summits Transl. Sci. Proc.* **2021**, *2021*, 410. [PubMed]
253. Hong, N.; Wen, A.; Stone, D.J.; Tsuji, S.; Kingsbury, P.R.; Rasmussen, L.V.; Pacheco, J.A.; Adekkanattu, P.; Wang, F.; Luo, Y.; et al. Developing a FHIR-based EHR Phenotyping Framework: A Case Study for Identification of Patients with Obesity and Multiple Comorbidities from Discharge Summaries. *J. Biomed. Inform.* **2019**, *99*, 103310. [CrossRef] [PubMed]
254. Yao, L.; Mao, C.; Luo, Y. Clinical Text Classification with Rule-based Features and Knowledge-guided Convolutional Neural Networks. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 31–39. [CrossRef] [PubMed]
255. Uzuner, O.; Solti, I.; Cadag, E. Extracting Medication Information from Clinical Text. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 514–518. [CrossRef] [PubMed]
256. Uzuner, O.; Solti, I.; Xia, F.; Cadag, E. Community Annotation Experiment for Ground Truth Generation for the I2B2 Medication Challenge. *J. Am. Med. Inform. Assoc.* **2010**, *17*, 519–523. [CrossRef] [PubMed]
257. Houssein, E.H.; Mohamed, R.E.; Ali, A.A. Heart Disease Risk Factors Detection from Electronic Health Records using Advanced NLP and Deep Learning Techniques. *Sci. Rep.* **2023**, *13*, 7173. [CrossRef]
258. Doan, S.; Collier, N.; Xu, H.; Duy, P.H.; Phuong, T.M. Recognition of Medication Information from Discharge Summaries using Ensembles of Classifiers. *BMC Med. Inform. Decis. Mak.* **2012**, *12*, 36. [CrossRef]
259. Fu, S.; Chen, D.; He, H.; Liu, S.; Moon, S.; Peterson, K.J.; Shen, F.; Wang, L.; Wang, Y.; Wen, A.; et al. Clinical Concept Extraction: A Methodology Review. *J. Biomed. Inform.* **2020**, *109*, 103526. [CrossRef] [PubMed]
260. Uzuner, O.; South, B.R.; Shen, S.; DuVall, S.L. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. *J. Am. Med. Inform. Assoc.* **2011**, *18*, 552–556. [CrossRef] [PubMed]

261. Naseem, U.; Thapa, S.; Zhang, Q.; Hu, L.; Masood, A.; Nasim, M. Reducing Knowledge Noise for Improved Semantic Analysis in Biomedical Natural Language Processing Applications. In Proceedings of the 5th Clinical Natural Language Processing Workshop, Toronto, ON, Canada, 14 July 2023; pp. 272–277.
262. Moscato, V.; Napolano, G.; Postiglione, M.; Sperlì, G. Multi-task Learning for Few-shot Biomedical Relation Extraction. *Artif. Intell. Rev.* **2023**, *56*, 13743–13763. [CrossRef]
263. Uzuner, O.; Bodnari, A.; Shen, S.; Forbush, T.; Pestian, J.; South, B.R. Evaluating the State of the Art in Coreference Resolution for Electronic Medical Records. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 786–791. [CrossRef] [PubMed]
264. Sun, W.; Rumshisky, A.; Uzuner, O. Evaluating Temporal Relations in Clinical Text: 2012 I2B2 Challenge. *J. Am. Med. Inform. Assoc.* **2013**, *20*, 806–813. [CrossRef] [PubMed]
265. Sun, W.; Rumshisky, A.; Uzuner, O. Annotating Temporal Information in Clinical Narratives. *J. Biomed. Inform.* **2013**, *46*, S5–S12. [CrossRef]
266. Kumar, V.; Stubbs, A.; Shaw, S.; Uzuner, O. Creation of a New Longitudinal Corpus of Clinical Narratives. *J. Biomed. Inform.* **2015**, *58*, S6–S10. [CrossRef]
267. Stubbs, A.; Uzuner, 0. Annotating Longitudinal Clinical Narratives for De-identification: The 2014 i2b2/UTHealth Corpus. *J. Biomed. Inform.* **2015**, *58*, S20–S29. [CrossRef] [PubMed]
268. Stubbs, A.; Kotfila, C.; Uzuner, O. Automated Systems for the De-identification of Longitudinal Clinical Narratives: Overview of 2014 I2B2/UTHealth Shared Task Track 1. *J. Biomed. Inform.* **2015**, *58*, S11–S19. [CrossRef] [PubMed]
269. Stubbs, A.; Filannino, M.; Soysal, E.; Henry, S.; Uzuner, O. Cohort Selection for Clinical Trials: N2C2 2018 Shared Task Track 1. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1163–1171. [CrossRef]
270. Henry, S.; Buchan, K.; Filannino, M.; Stubbs, A.; Uzuner, O. 2018 N2C2 Shared Task on Adverse Drug Events and Medication Extraction in Electronic Health Records. *J. Am. Med. Inform. Assoc.* **2019**, *27*, 3–12. [CrossRef] [PubMed]
271. McCreery, C.H.; Katariya, N.; Kannan, A.; Chablani, M.; Amatriain, X. Effective Transfer Learning for Identifying Similar Questions: Matching User Questions to COVID-19 FAQs. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual Event, 23–27 August 2020; pp. 3458–3465.
272. Soni, S.; Roberts, K. A Paraphrase Generation System for EHR Question Answering. In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; pp. 20–29.
273. Lau, J.J.; Gayen, S.; Ben Abacha, A.; Demner-Fushman, D. A dataset of Clinically Generated Visual Questions and Answers about Radiology Images. *Sci. Data* **2018**, *5*, 180251. [CrossRef] [PubMed]
274. He, X.; Zhang, Y.; Mou, L.; Xing, E.; Xie, P. PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv* **2020**, arXiv:2003.10286.
275. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.; Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. In Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 7 November 2019; pp. 2567–2577.
276. Hasan, S.A.; Ling, Y.; Farri, O.; Liu, J.; Müller, H.; Lungren, M. Overview of Imageclef 2018 MMedical Domain Visual Question Answering Task. In Proceedings of the CLEF Conference and Labs of the Evaluation Forum-Working Notes, Avignon, France, 10–14 September 2018.
277. Ben Abacha, A.; Hasan, S.A.; Datla, V.V.; Demner-Fushman, D.; Müller, H. Vqa-med: Overview of the Medical Visual Question Answering Task at Imageclef 2019. In Proceedings of the CLEF Conference and Labs of the Evaluation Forum-Working Notes, Lugano, Switzerland, 9–12 September 2019.
278. Ben Abacha, A.; Sarrouti, M.; Demner-Fushman, D.; Hasan, S.A.; Müller, H. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In Proceedings of the CLEF Conference and Labs of the Evaluation Forum-Working Notes, Bucharest, Romania, 21–24 September 2021.
279. Kovaleva, O.; Shivade, C.; Kashyap, S.; Kanjaria, K.; Wu, J.; Ballah, D.; Coy, A.; Karargyris, A.; Guo, Y.; Beymer, D.B.; et al. Towards Visual Dialog for Radiology. In Proceedings of the 19th SIGBioMed. Workshop on Biomedical Language Processing, Online, 9 July 2020; pp. 60–69.
280. Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; Germanidis, A. Structure and Content-guided Video Synthesis with Diffusion Models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7346–7356.
281. Jumper, J.M.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
282. Jiang, Y.; Gupta, A.; Zhang, Z.; Wang, G.; Dou, Y.; Chen, Y.; Fei-Fei, L.; Anandkumar, A.; Zhu, Y.; Fan, L. VIMA: General Robot Manipulation with Multimodal Prompts. *arXiv* **2023**, arXiv:2210.03094.
283. Jeblick, K.; Schachtner, B.; Dexl, J.; Mittermeier, A.; Stüber, A.T.; Topalis, J.; Weber, T.; Wesp, P.; Sabel, B.O.; Ricke, J.; et al. ChatGPT makes Medicine easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. *Eur. Radiol.* **2023**, 1–9. [CrossRef] [PubMed]
284. Wornow, M.; Xu, Y.; Thapa, R.; Patel, B.; Steinberg, E.; Fleming, S.; Pfeffer, M.A.; Fries, J.; Shah, N.H. The Shaky Foundations of Large Language Models and Foundation Models for Electronic Health Records. *NPJ Digit. Med.* **2023**, *6*, 135. [CrossRef] [PubMed]
285. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. *J. Priv. Confidentiality* **2016**, *7*, 17–51. [CrossRef]

286. Kerrigan, G.; Slack, D.; Tuyls, J. Differentially Private Language Models Benefit from Public Pre-training. In Proceedings of the Second Workshop on Privacy in NLP, Online, 16–20 November 2020; pp. 39–45.
287. Röösli, E.; Bozkurt, S.; Hernandez-Boussard, T. Peeking into a black box, the fairness and generalizability of a MIMIC-III benchmarking model. *Sci. Data* **2022**, *9*, 24. [CrossRef] [PubMed]
288. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
289. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef] [PubMed]
290. Wen, Y.; Wang, Z.; Sun, J. MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models. *arXiv* **2023**, arXiv:2308.09729.
291. Zack, T.; Lehman, E.; Suzgun, M.; Rodriguez, J.A.; Celi, L.A.; Gichoya, J.; Jurafsky, D.; Szolovits, P.; Bates, D.W.; Abdulnour, R.E.E.; et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. *Lancet Digit. Health* **2024**, *6*, e12–e22. [CrossRef] [PubMed]
292. Feng, Q.; Du, M.; Zou, N.; Hu, X. Fair Machine Learning in Healthcare: A Review. *arXiv* **2024**, arXiv:2206.14397.
293. Kumar, A.; Agarwal, C.; Srinivas, S.; Li, A.J.; Feizi, S.; Lakkaraju, H. Certifying LLM Safety against Adversarial Prompting. *arXiv* **2024**, arXiv:2309.02705.
294. Yuan, J.; Tang, R.; Jiang, X.; Hu, X. Large language models for healthcare data augmentation: An example on patient-trial matching. *AMIA Annu. Symp. Proc.* **2023**, *2023*, 1324. [PubMed]
295. Lai, T.; Shi, Y.; Du, Z.; Wu, J.; Fu, K.; Dou, Y.; Wang, Z. Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. *arXiv* **2023**, arXiv:2307.11991.
296. Kim, K.; Oh, Y.; Park, S.; Byun, H.K.; Kim, J.S.; Kim, Y.B.; Ye, J.C. RO-LLaMA: Generalist LLM for Radiation Oncology via Noise Augmentation and Consistency Regularization. *arXiv* **2023**, arXiv:2311.15876.
297. Allenspach, S.; Hiss, J.A.; Schneider, G. Neural multi-task learning in drug design. *Nat. Mach. Intell.* **2024**, *6*, 124–137. [CrossRef]
298. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
299. Zhang, Y.; Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.* **2018**, *5*, 30–43. [CrossRef]

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