Tampere University

Pinja Karttunen

LARGE LANGUAGE MODELS IN HEALTHCARE DECISION SUPPORT

Bachelor’s thesis
Faculty of Medicine and Health
Technology
Examiner: Mark van Gils
August 2023

ABSTRACT

Pinja Karttunen: Large Language Models in Healthcare Decision Support
Bachelor's thesis
Tampere University
Biotechnology and biomedical engineering
August 2023

Large language models (LLMs) have recently garnered significant attention due to their remarkable ability to assimilate vast amounts of information and effectively process natural language. In healthcare, natural language constitutes a substantial portion of medical data, rendering LLMs highly promising for various healthcare applications. This study seeks to explore the potential of LLMs in healthcare and clinical decision support (CDS), following PRISMA guidelines for reviews.

The analysis encompasses 44 LLMs, each influenced by several factors impacting their performance. Notably, the datasets utilized for pretraining and fine-tuning processes play a crucial role in determining the model's domain specificity. Furthermore, distinct model architectures are tailored for specific tasks, while prompting strategies are frequently employed to refine and enhance the model's performance.

LLMs exhibit considerable promise for a wide array of healthcare applications. For instance, LLMs possess the potential to efficiently handle and analyse medical information, facilitate contextual understanding among clinicians and patients, as well as automating the documentation of clinical notes and reports. Presently, however, their implementation within the field remains limited.

Notable improvements have been witnessed in the performance of current healthcare-oriented LLMs, with some achieving expert-level competence in medical question-answering (MQA). However, these LLMs face prominent challenges, encompassing ethical concerns, issues related to accountability, and a lack of appropriate regulations.

Nevertheless, this study reveals numerous promising applications in healthcare where LLMs could significantly augment the efficiency, accessibility, and manageability of healthcare delivery. Addressing the challenges LLMs encounter is essential for their seamless integration into practical healthcare applications. As a relatively new technology, the development of LLMs is still in its early stages, but their potential is evident through this study. Consequently, fostering collaboration among healthcare professionals, developers, regulators, and other stakeholders is imperative to cultivate dependable LLMs that align with the demands of the healthcare sector.

Keywords: large language model, clinical decision support, chatbot, performance, healthcare

TIIVISTELMÄ

Pinja Karttunen: Suuret kielimallit terveydenhuollon päätöksenteon tukena
Kandidaatin tutkielma
Tampereen yliopisto
Bioteknologia ja biolääketieteen tekniikka
Elokuu 2023

Suuret kielimallit ovat keränneet valtavasti huomiota lähiaikoina, sillä ne ovat osoittautuneet kyvykkäiksi ymmärtämään ja prosessoimaan suuria määriä tietoa sekä tuottamaan luonnollista kieltä ihmisen kaltaisesti. Terveydenhuollossa luonnollisen kielen käyttö on merkittävässä roolissa lääketieteellisessä datassa, mikä tekee suurista kielimalleista lupaavia teknologioita terveydenhuollon sovelluksissa. Tämä kirjallisuuskatsaus on toteutettu noudattaen PRISMA-ohjeistusta ja sen tavoitteena on tutkia suurten kielimallien potentiaalia terveydenhuollossa sekä niiden käyttöä kliinisessä päätöksenteossa.

Tässä tutkielmassa tarkastellaan 44:ää kielimallia, joiden suorituskykyyn vaikuttavat useat tekijät. Mallin esikoulutusdata ja hienosäätödata määrittelevät sen soveltuvuuden tietylle toimialueelle. Lisäksi erilaiset kielimallien arkkitehtuurit on suunniteltu erityisesti tiettyihin tehtäviin, ja syötteiden järjestelmällistä suunnittelua hyödynnetään usein tavoiteltujen tulosten saavuttamiseksi.

Suurilla kielimalleilla on useita käyttömahdollisuuksia terveydenhuollon sovelluksissa. Ne voivat esimerkiksi tehokkaasti käsitellä ja analysoida lääketieteellistä tietoa, helpottaa kliinikoiden ja potilaiden välistä informaation ymmärtämistä sekä automatisoida lääkärinlausuntojen ja muiden dokumenttien laatimista. Toistaiseksi suuria kielimalleja on kuitenkin hyödynnetty vielä melko vähän käytännön sovelluksissa terveydenhuollossa.

Viime vuosina suurten kielimallien suorituskyky on kehittynyt huomattavasti. Jotkut mallit ovat jopa saavuttaneet asiantuntijoiden tason vastatessaan lääketieteellisiin kysymyksiin. Kuitenkin nämä mallit kohtaavat myös merkittäviä haasteita, kuten eettisiä ongelmia, vastuullisuuskysymyksiä ja tarvittavien sääntelyjen puuttumista.

Tämä tutkimus esittelee lukuisia lupaavia terveydenhuollon käyttökohteita suurille kielimalleille, jotka voisivat olennaisesti parantaa terveydenhuollon tehokkuutta, saavutettavuutta ja hallittavuutta. Jotta näitä kielimalleja voitaisiin laajasti hyödyntää terveydenhuollon sovelluksissa tulevaisuudessa, on tärkeää käsitellä niiden kohtaamia haasteita. Vaikka suuret kielimallit ovat suhteellisen uusi teknologia ja niiden kehitys on edelleen alkuvaiheessa, tämä tutkimus osoittaa niiden lupaavat mahdollisuudet mullistaa terveydenhuolto. Seuraavaksi terveydenhuollon ammattilaisten, mallien kehittäjien, virkamiesten sekä muiden sidosryhmien tulisi tehdä yhteistyötä luotettavien mallien kehittämiseksi, jotka vastaavat alan vaatimuksia.

Avainsanat: suuri kielimalli, kliinisen päätöksenteon tuki, keskustelubotti, suorituskyky, terveydenhuolto

CONTENTS

1. INTRODUCTION 1

2. METHODOLOGY.. 3

2.1 Search strategy and study selection.. 3

2.2 Data extraction and analysis 4

3. THEORETICAL BACKGROUND.. 5

3.1 Large language models in healthcare 7

3.2 Prompting strategies 10

3.3 Definition of performance 11

4. LARGE LANGUAGE MODEL APPLICATIONS IN HEALTHCARE AND DISEASE DIAGNOSTICS 12

4.1 Potential use cases 12

4.1.1 Clinical workflow 12

4.1.2 Other possibilities.. 15

4.2 Current applications 16

4.3 Performance of current models 19

4.3.1 Diagnostics and triage 19

4.3.2 Radiology.. 19

4.3.3 Benchmarks.. 20

4.3.4 Others... 23

5. CHALLENGES OF LARGE LANGUAGE MODELS IN HEALTHCARE AND DISEASE DIAGNOSTICS... 28

6. DISCUSSION... 33

7. CONCLUSIONS... 37

REFERENCES.. 39

LIST OF ABBREVIATIONS AND MARKINGS

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
CAD	Computer-Aided Diagnosis
CDS	Clinical Decision Support
ChatGPT	Chat Generative Pre-trained Transformer
CoT	Chain-of-Thought
EHR	Electronic Health Record
EMR	Electronic Medical Record
ER	Ensemble Refinement
FDA	U.S. Food and Drug Administration
GPT	General Pre-trained Transformer
LLM	Large Language Model
ML	Machine Learning
MQA	Medical Question Answering
NER	Named Entity Recognition
NLI	Natural Language Inference
NLP	Natural Language Processing
NLU	Natural Language Understanding
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QA	Question Answering
RE	Relation Extraction
RNN	Recurrent Neural Network
SC	Self-Consistency
SOTA	State-Of-The-Art
SQ	Self-Questioning
STS	Semantic Textual Similarity

1. INTRODUCTION

LLMs have garnered significant attention recently due to their impressive performance in various natural language understanding tasks (NLU), such as question-answering (QA), machine translation, text summarization and sentiment analysis [1]. In the context of healthcare delivery, LLMs have the potential to play a crucial role, facilitating interactions between healthcare providers, researchers, and patients [2]. These models have been described as the most remarkable achievements in the field of artificial intelligence (AI), as they possess the capacity to absorb vast amounts of information and interact in a manner resembling human communication [3]. Nevertheless, the extent of LLMs' performance in the healthcare domain remains uncertain – do they meet expectations, or do they fall short of satisfying the specific needs and limitations of healthcare?

For decades, the field of AI has experienced waves of heightened excitement, followed by disappointments and setbacks, leading to slow progress. The transformer model architecture, introduced by Google in 2017, served as a foundation for more advanced LLMs capable of containing trillions of parameters [4]. Consequently, various LLMs emerged such as the Generative Pretrained Transformers (GPT) series by OpenAI and the Bidirectional Encoder Representative from Transformers (BERT) by Google. These models, often referred to as foundation models, laid the groundwork for other LLM variants like ChatGPT and BioBERT. LLMs find prominent use in conversational AI, commonly referred to as "chatbots". Notably, the release of ChatGPT chatbot in November 2022 marked a noteworthy turning point for LLMs, attracting over 100 million users exploring the capabilities of LLMs within two months of its launch [5].

In the context of healthcare, LLMs have exhibited potential in assisting clinicians with administrative tasks, medical text summarization, data analysis, and optimizing CDS. Additionally, LLM chatbots can serve as personalized assistants for patients, providing health guidance and support. LLMs also offer translation capabilities for medical texts and can be valuable in educational and examination contexts. However, there are notable concerns, including privacy issues, bias amplification, tendency to produce hallucinations or false information, and ethical considerations. This thesis aims to address both the advantages and challenges associated with the application of LLMs in healthcare.

Given the rapid advancements in LLM technology over the past few years, this review focuses on literature from 2020, primarily centred on the use of LLMs in CDS for healthcare. This review first introduces the methodology employed, followed by addressing theoretical aspects. Subsequently, this study explores the current and potential applications of LLMs in CDS, reports the performance of existing models, discusses prevailing challenges, and considers future prospects.

This study aims to thoroughly investigate the potential of LLMs in healthcare and CDS systems. Through an examination of the current literature, this review endeavours to determine whether LLMs can effectively overcome the challenges of implementation in the healthcare domain. To achieve these objectives, this review analyses the suitability of current LLMs, considering essential aspects such as performance, domain specificity, model architecture, and prompting strategies, and their corresponding influence on model performance. By comprehensively analysing the current state and challenges faced by LLMs in healthcare, this study seeks to discover their potential to revolutionize healthcare practices in the future. Furthermore, the aim is to provide valuable insights into the necessary actions that different stakeholders must undertake to facilitate the practical utilization of LLMs.

2. METHODOLOGY

This literature review was conducted by following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [6]. The PRISMA analysis facilitated the organization of the material by clustering the articles according to the relevant keywords. The data were then categorized into mind maps centred around key topics.

2.1 Search strategy and study selection

Adhering to the PRISMA methodology ensured transparent collection and systematic evaluation of the selected reports for this review. The research questions were defined, and the topic was narrowed down to establish the inclusion criteria.

Suitable reports were searched from databases including PubMed, Andor and Google Scholar. The selected search terms comprised “large language model*”, “disease*”, “diagnostic*”, “future”, “challenge*”, and “bias*”. The asterisk (*) symbol was used for truncation, indicating that any characters could appear after the specified term. Boolean operators such as “AND” and “OR” were employed to combine the terms appropriately. Specific parameters were set to narrow down the search results, focusing on medical research and the publications started from 2020.

After identifying the records, a screening process was performed, and certain studies were excluded based on predetermined criteria. The exclusion criteria encompassed records, that fell outside the scope of the study. Additionally, records that extensively discussed the operational principles of the large language model or provided a generalized coverage of covered artificial intelligence or natural language processing (NLP) were excluded. Furthermore, records presenting the usage of ChatGPT in medical applications without further deliberation were eliminated.

From the aforementioned databases, 66 records were identified, of which six duplicates were removed. The abstracts and conclusions of these records were observed, resulting in the exclusion of nine reports. Furthermore, citation search 51 records were identified via citation searching and three from websites, which resulted in total of 109 reports for full-text screening. After evaluating the full-texts, 46 reports were rejected, mainly due to content repetition with other included literature. Finally, 70 reports met the eligibility criteria for the final review. Figure 1 illustrates the record selection process in the form of a PRISMA flowchart.

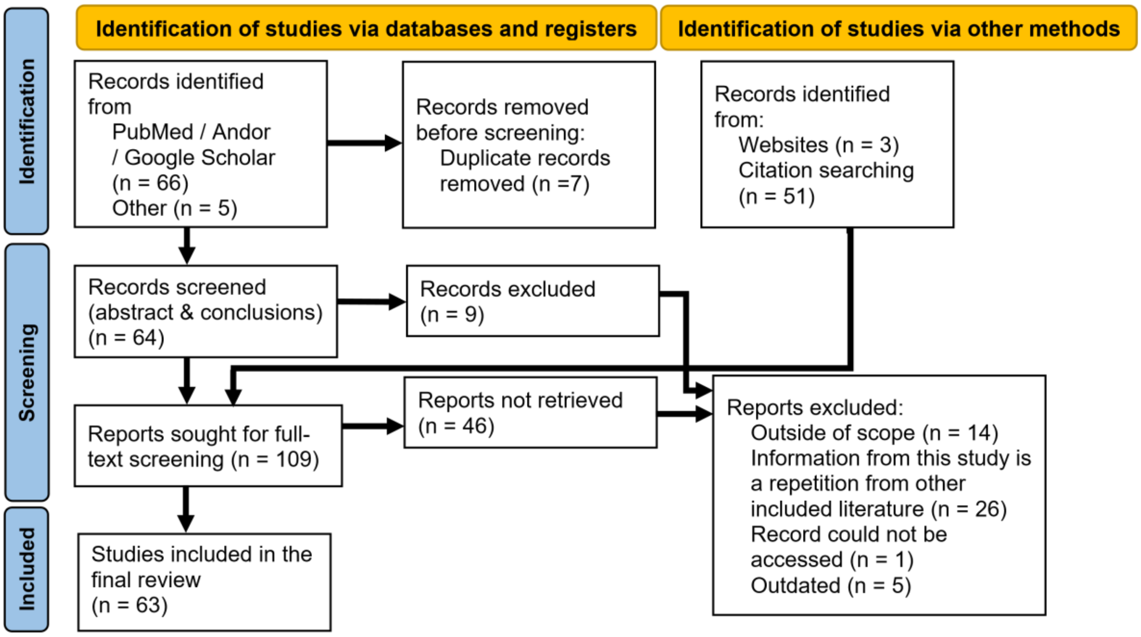

Identification
Screening
Included
Identification of studies via databases and registers
Identification of studies via other methods
Records identified from PubMed / Andor / Google Scholar (n = 66) Other (n = 5)
Records removed before screening: Duplicate records removed (n =7)
Records identified from: Websites (n = 3) Citation searching (n = 51)
Records screened (abstract & conclusions) (n = 64)
Records excluded (n = 9)
Reports sought for full-text screening (n = 109)
Reports not retrieved (n = 46)
Reports excluded: Outside of scope (n = 14) Information from this study is a repetition from other included literature (n = 26) Record could not be accessed (n = 1) Outdated (n = 5)
Studies included in the final review (n = 63)

Figure 1: PRISMA flowchart. [6]

2.2 Data extraction and analysis

The material was arranged by grouping the articles based on the keywords, facilitating a systematic approach that provided insight into the available information on each topic. Keywords were identified manually from each article and for each keyword, the sources in which the keyword was used were documented. Using this method, the extracted information was recorded, and the clustered material proved advantageous for analysis, as it facilitated the identification of articles relevant to the research objectives. Additionally, full-article screening revealed recurring topics, around which mind maps were constructed to further analyse the studies. This technique aided in visualizing the relationships between similar findings.

The Zotero software was employed to store and organize the collected records [7]. This tool automatically identified some keywords from each article, thus establishing a basis for keyword clustering. As the research progressed, notes were appended to Zotero, thereby augmenting the methodical organization of the articles. Furthermore, the capability to relocate excluded articles to a trash folder without irreversible deletion provided a practical solution. Notably, Zotero generated accurate citations for the sources, offering a time-saving advantage in the citation management process.

3. THEORETICAL BACKGROUND

The term artificial intelligence refers to computer systems that aim to achieve the capability of learning, reasoning, and problem-solving in a way similar to that of humans [8]. When combined with vast amounts of data and powerful computing resources, machine learning (ML) comes into play. ML, a subfield of AI, functions without the need for explicit programming. The principle involves using training data as input to identify patterns and train a predictive model. Once trained, the model can make predictions by analysing new, unseen data, continuously improving its performance. [8] ML can be classified into supervised learning, unsupervised learning, and reinforcement learning [9].

Deep learning utilizes algorithms that operate at multiple layers of abstraction to identify complex patterns from input data. These layers progressively transform raw data into novel and more abstract presentations. The advantage of these non-linear operations lies in their ability to learn highly intricate functions, and these algorithms may be supervised or unsupervised. [8,10,11] Recurrent neural network (RNN), convolutional neural network (CNN), and deep reinforcement learning are deep learning architectures and algorithms that play significant roles in NLP [9,12]. NLP focuses on analysing and representing human languages for tasks such as speech recognition, machine translation, text generation, QA, and information extraction [8]. Figure 2 illustrates the relationships of the presented concepts.

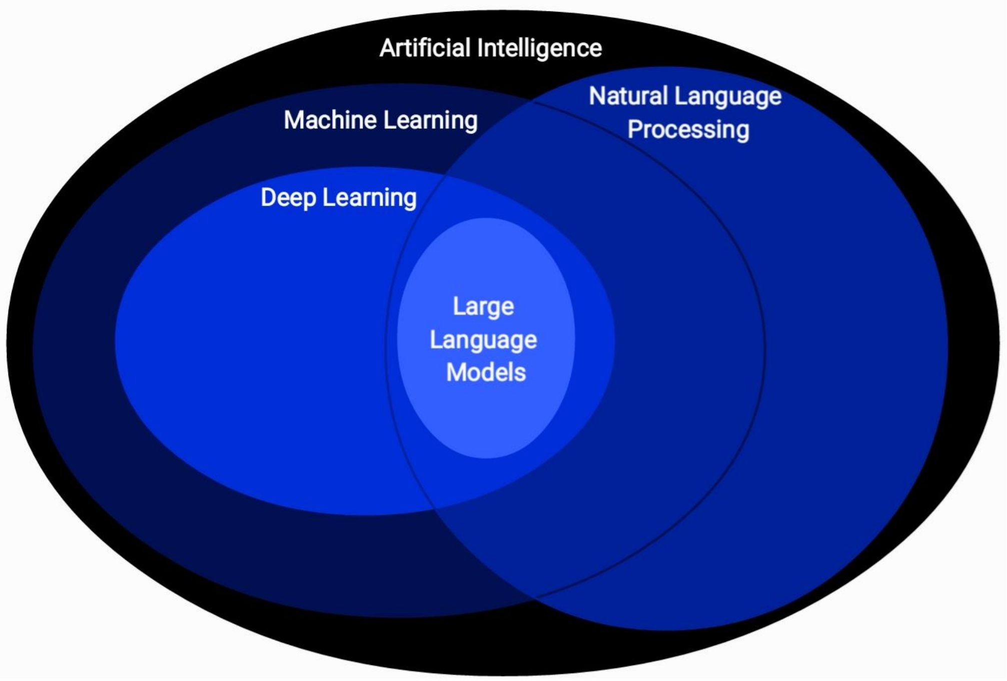


Figure 2: The intersections of AI, ML, NLP, DL, and LLMs

Transformer models, the most recent architecture models in NLP, are typically trained using self-supervised learning and fine-tuning [12]. In self-supervised learning, the model first learns about the specific medical field without explicit labels. The model leverages vast amounts of unlabelled data to learn complex structures and features. Following this preliminary training, the models are further trained on a smaller labelled dataset specific to the medical domains. This fine-tuning process enables the model to connect its preliminary knowledge with the explicitly labelled dataset, enhancing its ability to solve its primary task. [13]

Figure 3 presents different training approaches for LLMs. The first row demonstrates pretraining of domain-specific language models from scratch, solely focusing on the clinical domain dataset. The second row showcases models trained using Domain-Adaptive Pretraining (DAPT), where a general-domain model is further pretrained with clinical domain's dataset. The third row depicts fine-tuning a general model's foundation, while the fourth row represents a general-purpose model. In-context learning can be applied to all models, involving guiding the model with various prompting strategies, as discussed in Section 3.2, to enhance their adaptability and performance in specific tasks. [14]

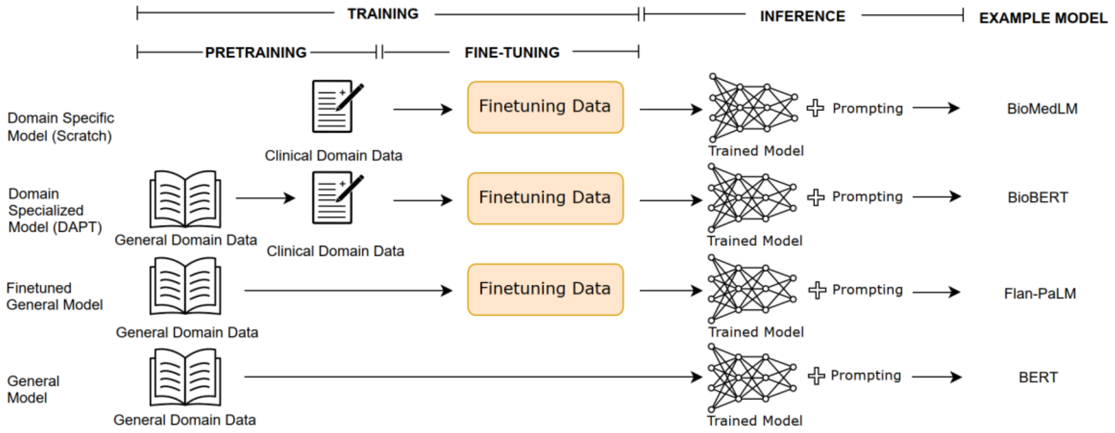


Figure 3: Training Approaches for LLMs, edited from Lehman et al. [14]

3.1 Large language models in healthcare

In recent years, there has been a tremendous growth in the development of LLMs, with continuous updates and improvements to existing models. Moreover, new variations of models such as GPT and BERT are being specifically developed to cater to specific requirements and demands. For instance, domain-specific LLMs like BioGPT and PubMedBERT have been developed to address the unique needs of the biomedical field and its applications. To adapt the model for the healthcare domain, it can be either pretrained using clinical domain data or fine-tuned from a general model.

LLMs can be classified into three distinct architectures: encoder-only, decoder-only, and encoder-decoder models, each differing in how they handle information between input and output. Encoder-only LLMs process and represent the input data by encoding the text into continuous vector representations, capturing contextual information in the process. In contrast, decoder-only models generate output based on the given context. They do this without explicitly encoding the input, often by predicting the next token in a sequence. Encoder-decoder models combine these approaches, first encoding the input data and then decoding it to generate output. [15] The original transformer model [4], which serves as the basis for subsequent LLMs, employed an encoder-decoder architecture.

Decoder-only models are typically used for text generation tasks, while encoder-only models find applications in classification, sentiment analysis, named entity recognition (NER), and other downstream tasks. Encoder-decoder models are employed in tasks where output is generated from the input, such as text translation and summarization. [15]

LLMs can be trained using large general datasets from diverse domains to provide a broad understanding of various subjects. Alternatively, they can be trained using domain-specific data, such as healthcare data, to offer more precise information. If a model is initially trained with general data and subsequently fine-tuned with domain-specific data, it is considered a general domain model. In this study, the model is classified as a specific domain model, if it is trained specifically for healthcare purposes without general domain data, as illustrated in the first row in figure 3.

This study focuses on LLMs developed by companies and institutions. Companies like Google, OpenAI, Microsoft, DeepMind, Meta AI, NVIDIA, Healx, and ZoomRx, as well as institutions such as Stanford University, Massachusetts Institute of Technology, University of Florida, National Centre for Biotechnology Information, MosaicML, Allen Institute for AI, King’s College London, and University College London, have contributed to the development of these LLMs.

Figure 4 illustrates the current healthcare LLMs reviewed in this literature study. The figure provides a timeline of LLM releases and the corresponding publishers. Encoder-only models are represented in green, decoder-only models in red, and encoder-decoder models in blue. General domain models and specific domain models trained are differentiated using lighter and darker colour, respectively. Additionally, the logos of companies are explicitly displayed in the figure, while models developed by institutions are represented by a unified icon.

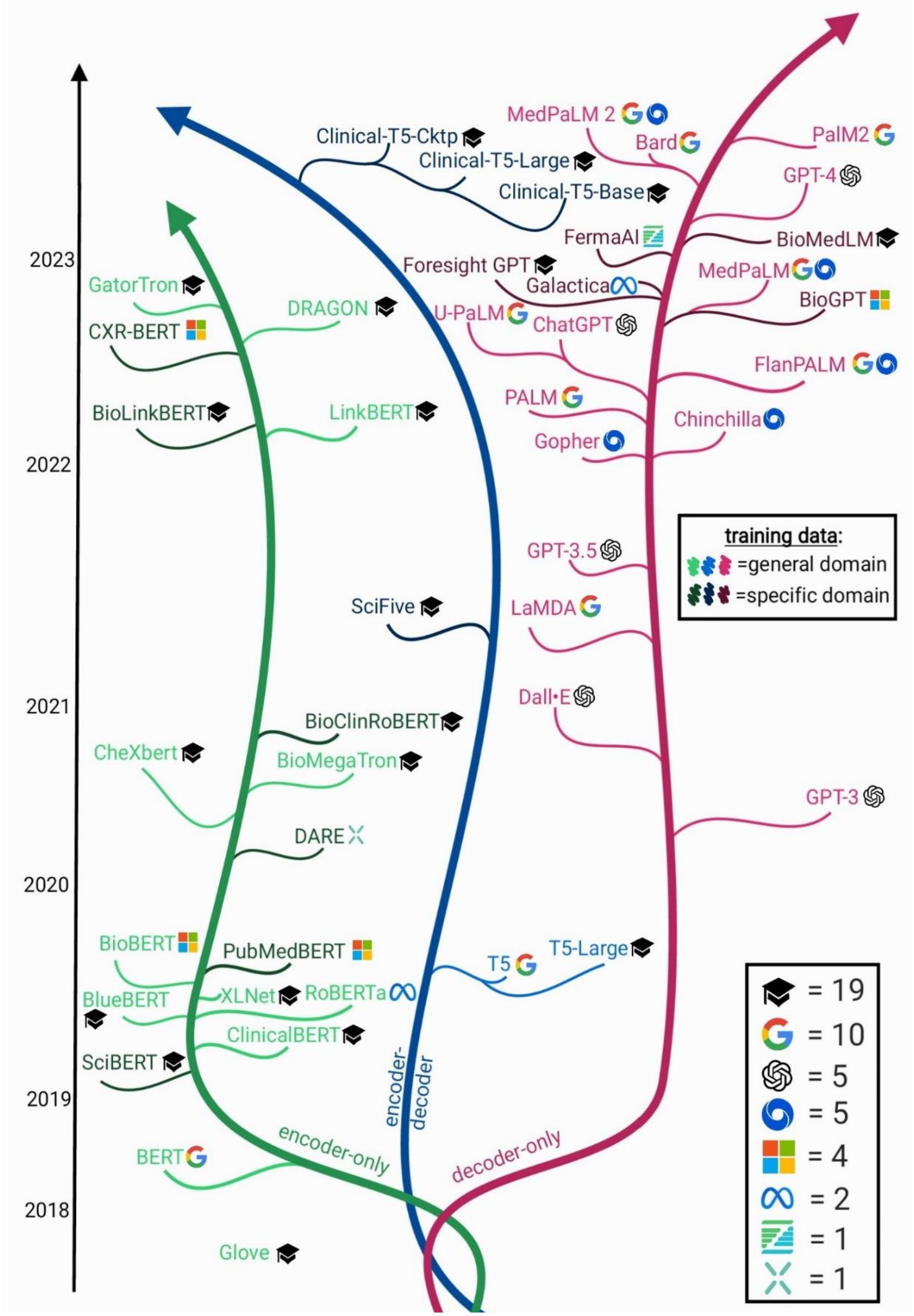


Figure 4: LLMs utilized in healthcare [1,12,14–44]

In total, there are 15 domain-specific models and 29 general domain models among the 44 models analysed in this study. 19 models have been developed by institutions, with Stanford University and Massachusetts Institute of Technology contributing 5 models from each. Google, OpenAI, DeepMind, and Microsoft being the most prominent contributors among the models developed by companies.

Figure 4 demonstrates that the release of OpenAI's GPT-3.5 model in March 2022 triggered a wave of model publications from other companies and institutes, some of which utilized the GPT-3.5 as a foundation. Recently, decoder-only models, have a major role compared to other model types. This can be attributed to their architecture, which allows them to efficiently capture long-range dependencies and learn contextual representations [15].

3.2 Prompting strategies

Prompting strategies are frequently employed to enhance and refine the performance of models without any fine-tuning or updates. Given the considerable cost associated with fine-tuning LLMs and the remarkable success of in-context learning, these strategies are widely adopted. In the standard prompting technique, a concise prompt is provided to the model to achieve the desired output. To further guide the model towards improved performance, more systematic prompting methods are employed. Utilizing defined prompting strategies, such as few-shot, zero-shot, chain-of-thought (CoT), self-consistency (SC), ensemble refinement (ER), self-questioning (SQ), allows for reliable performance comparison. [1,2,23]

In zero-shot prompting, the model is expected to perform without exposure to any labelled dataset. Conversely, in few-shot prompting, the model is given a few examples of input-output pairs for adaption. CoT involves providing a step-by-step explanation towards the final answer, guiding the model through logical reasoning. [1,23]

Wang et al. introduced a strategy inspired by human-like reasoning, called self-questioning prompting. SQ aims to deepen the model's comprehension of the desired concepts by posing multiple targeted questions about the task. These questions are designed to cover different aspects of the key elements, thereby enhancing the model's ability to provide improved answers. [1]

In tasks that involve complex deductions, there may be multiple paths to arrive at the same answer. Therefore, it is beneficial to elicit multiple outputs from the model and select the final answer based on the majority vote. This prompting strategy is known as self-consistency. [2]

Singhal et al. devised the ER strategy, which combines the characteristics of CoT and SC. Firstly, the model is provided with CoT prompt and a question resulting in the generation of multiple reasoning paths towards various answers. Secondly, the model is prompted to produce a refined explanation and answer based on the generations from the previous step, the original prompt, and question. The second stage of ER, which can be interpreted as SC, is typically repeated multiple times, and the final answer is determined through a majority vote. [23]

3.3 Definition of performance

Considerable research has been done to evaluate the performance of current LLMs and compare different models. Performance assessment often relies on standardized benchmark datasets that encompass various forms of MQA [2]. Furthermore, biomedical benchmarks are used in medical licensing examinations [45]. Benchmarks involve a range of question types, including both long, narrative answers and shorter, multiple-choice questions. Each dataset focuses on distinct aspects of medical knowledge, such as medical exams, medical research, or consumer health. Benchmarks can be further divided into closed domain or open domain datasets. In closed domain datasets, answers are restricted to a predetermined set of sources, while open domain datasets have no such limitations. [2]

For example, the MedQA dataset contains multiple-choice questions sourced from the US medical licencing exam (USMLE), and MedMCQA dataset is collected from Indian medical school entrance exams (AIIMS and NEET-PG) [2,46]. PubMedQA focuses on multiple-choice questions derived from biomedical scientific literature. On the other hand, MedicationQA provides long-form answers that cater to general medical knowledge sought by consumers. [2] Another dataset, MedSTS, comprises sentence pairs annotated from clinical notes and can be utilized for evaluating the semantic textual similarity of two texts [12].

Using benchmark datasets, the performance of LLMs can be evaluated on diverse clinical language understanding tasks. In addition to straightforward QA, and semantic textual similarity (STS) task, a natural language inference (NLI) task intends to verify whether a conclusion can be inferred from text in question [12]. Relation extraction (RE) involves jointing and classifying the relationships and entities from text. If the intention is to classify the document into predefined categories, the task is called document classification. [20] Alternatively, if classification is performed according to predetermined entities, an NER task is in question [1].

4. LARGE LANGUAGE MODEL APPLICATIONS IN HEALTHCARE AND DISEASE DIAGNOSTICS

LLMs possess advanced capabilities in processing and generating natural language, rendering them highly suitable for a wide range of healthcare applications. These applications encompass the handling and analysis of medical information, as well as the facilitation of contextual understanding for healthcare professionals and patients. This chapter delves into the utilization of LLMs in healthcare, exploring various potential use cases and existing applications in the field.

4.1 Potential use cases

Electronic health records (EHRs) and electronic medical records (EMRs) are both repositories of medical information pertaining to patients. EMRs are typically confined to a specific healthcare organization, whereas EHRs cover information related to encounters with multiple components of the healthcare system throughout a person’s life, and can possibly be shared and accessed by multiple healthcare organisations [47]. For the sake of simplicity, this section employs the term EHR to encompass scenarios applicable to both EHRs and EMRs.

4.1.1 Clinical workflow

Medical information is expanding rapidly, posing various opportunities but also significant challenges for healthcare professionals. The retrieval process needs to be credible, relevant, accessible, fast, and user-friendly. EHRs contain extensive medical information about patients, including structured and unstructured data elements such as medical history, medications, diagnoses, and test results. [45,48] LLMs have the potential to assist healthcare professionals in managing medical literature, interpreting patient records, and developing personalized treatment plans [1].

Preventive care and overall health maintenance are crucial in disease prevention. Smartphone applications and wearable technology (wearables) offer solutions for promoting people's health. LLMs can **analyse** personal health data from these modern applications. Moreover, integrating chatbots with these applications can provide **health**

guidance. For example, a chatbot can provide medication information including potential interactions with other substances and possible side-effects. Additionally, chatbots can serve as **virtual assistants** for evaluating the need for care, managing health information, scheduling appointments, or providing pre-operative instructions. [49]

Efficient medical **triaging** of patients requires fast reasoning and ability to determine the severity and urgency of their conditions [49]. LLMs, capable of managing data from multiple sources, can utilize patient pre-test odds, diagnostic likelihood ratios, and EHRs to guide triage decisions and improve efficiency [50].

Clinicians often need to review a patient’s medical history before an appointment. LLMs can **summarize** patients’ EHRs, saving clinicians time by extracting meaningful information such as symptoms, diagnoses, treatments, imaging reports, and lab results from the record. Clinicians can also pose specific questions to LLMs, leveraging the information within EHRs. [49,51] SNOMED CT is the world’s most extensive clinical terminology, which determines the global standards for clinical terminology [52]. Patient records can be inputted into standardised ontology based LLMs, such as foresight GPT, to generate probabilistic **forecasts** [45]. Additionally, LLMs can summarize relevant, patient-specific information from scientific research and medical papers to quickly identify key findings and insights for clinical decisions [53,54].

A **CDS** optimizes clinical decision-making by providing information and recommendations to physicians, patients, and other stakeholders [51,55]. The adoption of EHR has increased the use of CDS to enhance healthcare services and patient outcomes, since CDS can provide for example treatment planning and diagnosis suggestions from EHRs [55]. Integrating LLMs to CDS infrastructure can improve the accuracy and efficiency of the systems [50]. Certified EHRs require rule-based and data-driven CDS **alerts** that provide task- and patient-specific recommendations, improving clinical quality and addressing disparities. These alerts include potential drug interactions, allergies, or other considerations that should be considered in decision-making. However, clinicians often experience alert fatigue due to the constant influx of alerts, resulting high rate of ignored or cancelled alerts. LLMs can improve CDS logic and optimize CDS alerts to mitigate these challenges. [55]

With more accessible imaging technologies and aging population, medical imaging volumes are predicted to increase [50]. LLMs hold potential for integration with **computer-aided diagnosis** (CAD) systems in medical imaging. By combining the medical knowledge and logical reasoning of LLMs with the vision understanding capabilities of CAD systems, the interpretability of results can be enhanced for clinicians and patients.

For example, medical images can be processed by CAD models, and the output can be translated into natural language. LLMs can **summarize** these results and facilitate discussions on symptoms, diagnosis, and treatment. However, LLMs currently face challenges in understanding visual information, limiting their support in clinical decision-making. The proposed strategy of Wang et al. for utilizing LLMs in CAD in Medical Imaging is illustrated in the Figure 5. [56]

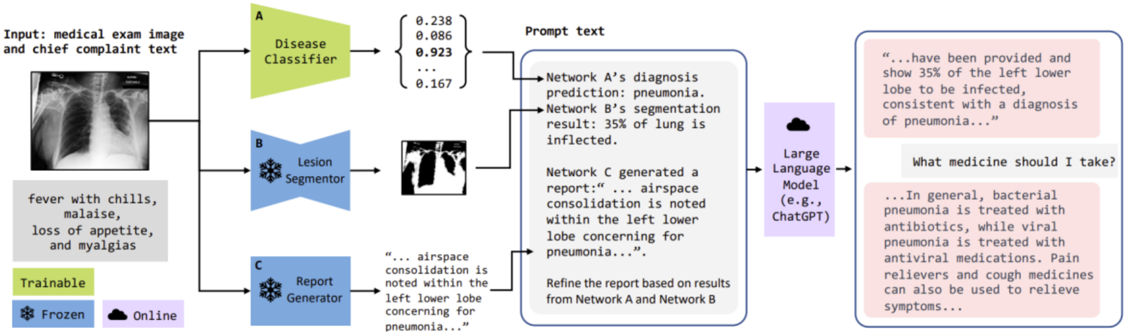

Input: medical exam image and chief complaint text
fever with chills, malaise, loss of appetite, and myalgias
Trainable
❄ Frozen
☁ Online
A
Disease Classifier
B
❄ Lesion Segmentor
C
❄ Report Generator
0.238
0.086
0.923
...
0.167
"... airspace consolidation is noted within the left lower lobe concerning for pneumonia..."
Prompt text
Network A's diagnosis prediction: pneumonia.
Network B's segmentation result: 35% of lung is inflected.
Network C generated a report:" ... airspace consolidation is noted within the left lower lobe concerning for pneumonia...".
Refine the report based on results from Network A and Network B
☁
Large Language Model (e.g., ChatGPT)
"...have been provided and show 35% of the left lower lobe to be infected, consistent with a diagnosis of pneumonia..."
What medicine should I take?
...In general, bacterial pneumonia is treated with antibiotics, while viral pneumonia is treated with antiviral medications. Pain relievers and cough medicines can also be used to relieve symptoms...

Figure 5: LLM utilization in CAD in Medical Imaging [56]

Documentation consumes a significant amount of physicians' and nurses' time [57]. Clinical notes include various elements such as summaries of admissions, medical history, consultation notes, and examination findings [58]. LLMs can generate draft documents for medical professionals to review and edit [59]. By leveraging LLMs for paperwork, clinicians can focus their attention on providing patient care [60]. LLMs specifically trained for certain field, such as radiology, can aid in creating report templates for specialists [27]. Furthermore, specific models can be trained to generate summaries directly from medical dialogues [61].

Clinical notes are typically written in highly technical language containing jargon, which can be challenging for patients to understand as they are traditionally not meant for patients. LLMs can **simplify** clinical documents into patient-friendly language, significantly improving the accessibility of medical information. [54,60] Additionally, LLMs can **translate** medical documents into a patient's native language, utilizing appropriate medical terminology, further enhancing availability and comprehension [62].

Chatbots can be implemented in follow-up and postoperative care to respond to patients' concerns in a timely manner. Chatbots can instantly **provide educational information** or care instructions tailored to the patient's needs. [49,63] Furthermore, these applications can offer emotional and psychological **support**, allowing patients to discuss personal difficulties without judgement. The ability to provide remote support from home can be highly valuable for numerous patients. [63]

Wearables play a convenient role in **remote patient monitoring**. The devices can be incorporated into health coaching programs to prevent or improve chronic conditions such as obesity or cancer. [63] As previously discussed, LLMs can analyse data from wearable devices, providing results to the patient or to healthcare professionals for intervention in potential setbacks [49]. Wearables can also include LLM-generated medication reminders to ensure patients receive the correct medication at the appropriate time and dosage [53]. Integrating these remote patient care systems with chatbots can harness the advantages of both approaches [63].

4.1.2 Other possibilities

LLMs have a great potential for use in **education** of healthcare students, patients, and clinicians [53,57,64]. In the context of learning, LLMs can generate mnemonics, explanations of concepts, and medical exams or questions to aid learners [64]. Moreover, LLMs can generate patient education materials to explain disease-specific concepts in a patient-friendly manner [53]. Additionally, LLMs can assist clinicians in efficient medical information searching, leveraging their powerful search engine capabilities [65]. These education materials can be translated into the recipient's preferred language by LLMs to ensure comprehensibility [66].

LLMs can also support the process of **writing medical reports** for physicians [64]. Medical reports are composed of the patient's clinical notes, which include background information, medical history, physical examination, specimens, treatments, and suggested opinions. These reports serve as a means of communication between physician and legal system, often required for criminal or civil transactions involving entities such as the police, government tribunals, insurance companies, lawyers, or patients themselves. [67] In social media posts on platforms like TikTok [68] and Twitter [69], Dr Cliff Stermer, a rheumatologist, showcased the capability of ChatGPT to draft a letter addressed to an insurance provider. The demonstration involved the drafting of a comprehensive letter that included references and was centred around a patient with systemic sclerosis who required approval for echocardiogram. These posts garnered considerable attention, generating widespread views, and sparking discussions around the topic.

For **examination** purposes, LLMs can enhance clinical trial matching and clinical trial enrichment processes through their abilities in clinical information extraction and clinical reasoning. [57,70]. LLMs can improve compatibility between the EHRs and eligibility criteria by identifying relevant ontologies and terminologies and selecting appropriate patients for trials [71]. Furthermore, LLMs can be utilized to identify patterns from research data, and generate charts based on them [53,72]. LLMs can also assist researchers in creating examination documents and translating articles and studies [53].

Several studies have explored the use of LLMs in **speech-related** applications. Amini et al. and Agbavor et al. conducted studies demonstrating the potential of LLMs in early detection of Alzheimer's disease and related dementias. Speech analysis can be a valuable predictor of cognitive impairments associated with Alzheimer's disease. Both studies employed similar approaches, utilizing speech-to-text conversion for voice recordings, and employing LLMs to detect cognitive impairments from generated text. [73,74] LLMs can also be utilized in other neurogenerative conditions that cause speech impairments. Willet et al. presented a study on decoding brain signals and attempted handwriting movements into real-time translated text, where LLMs were applied to autocorrect errors in the generated output. [75] Figure 6 illustrates some of the potential use cases of LLMs. [76]

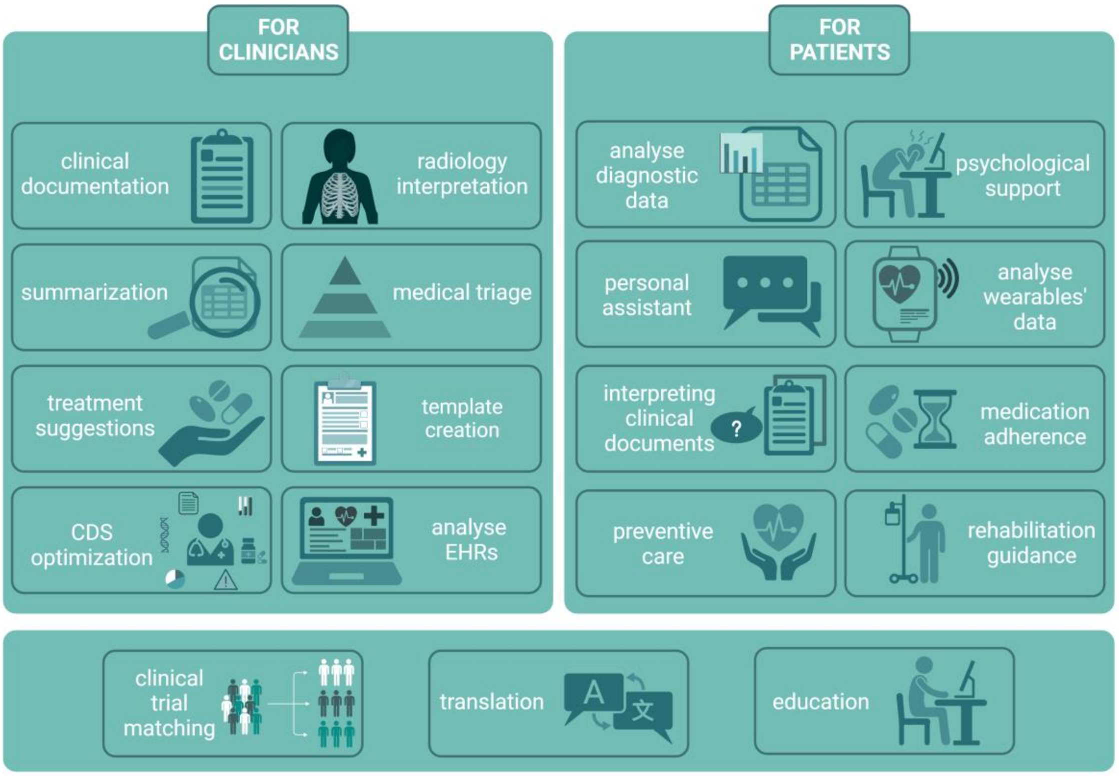


Figure 6: Potential use cases of LLMs. Adapted from Meskó and Topol [76]. Created with BioRender.com

4.2 Current applications

To assist physicians with their **documentation** tasks, Nabla Copilot, a GPT-3 based digital assistant accessed as a Chrome extension, proves beneficial [77]. Copilot transcribes and repurposes information from patient encounters into prescriptions, referrals, follow-up appointment letters, and consultation summaries. The clinical notes can be edited and further updated in patients' EMRs. The application provides a reliable tool to accurately capture every word from encounters, allowing physicians to shift their focus to the consultation with the patient. [77,78]

Nuance, a leading provider of clinical speech recognition products for healthcare providers, is widely used in the U.S. with a significant market share among radiologists and other physicians [79]. In March 2023, Nuance introduced Dragon Ambient eXperience (DAX) Express, their latest **documentation** product, which leverages the revolutionary capabilities of the GPT-4 model to enhance their previous market-leading DAX model. The integration of the generative AI model automates the drafting of clinical notes, saving valuable time for clinicians. [80]

Doximity, in collaboration with ChatGPT, has developed a beta version called DocsGPT on their digital platform to assist physicians with **administrative tasks**. This product has been fine-tuned in collaboration with doctors to ensure accurate adaptation to the healthcare field. DocsGPT allows for the drafting and faxing pre-authorization and appeal letters to insurers digitally. As previously mentioned, LLMs cannot generate documents without doctors' review, so the product prompts the user to ensure accuracy before submission. Doximity's future goal is to further develop DocsGPT for commercialization and broaden its functionalities. [81]

One of the fastest-growing health service provider Babylon Health combines LLM in their AI-driven primary care **chat** service. LLM is used to assist healthcare provider, for example asking by questions about the patient's symptoms and delivering personalized medical advice. [82]

MedMatch Network serves as a platform for facilitating patient referral management and secure information exchange, connecting physicians, other service providers, and patients through over 1.7 million profiles. It integrates EHR systems, automates appointment scheduling, and provides updates and alerts to both providers and patients. MedMatch Network utilizes ChatGPT to offer a **chatbot** for MQA to assist patients. [83]

The managed **chatbot** service ChatBeacon offers emotional assistance and support, particularly for individuals with mental health problems. It aims to empower individuals by providing self-care exercises and crisis support. The chatbot operates 24/7 and delivers individually tailored responses, making it valuable for those in need of non-judgemental support. [84]

Livewello and Amazfit incorporate **chatbot** features in their respective applications [85]. Livewello is a genome data-analysis tool that incorporates GeneChat, enabling users to ask questions about their genetic data or their personal health record [86]. Amazfit manufactures health management wearables that offer LLM-powered chat for personal coaching [87].

Ferma is a **chatbot** specifically designed to answer life science questions. It utilizes ChatGPT along with specific pharma and clinical datasets for question-answering. What sets Ferma apart is that it provides step-by-step reasoning and used sources for the user. [39]

Bionic Health has incorporated GPT-4 capabilities into their preventative care platform, which aims to connect patients to physicians and optimize their health between in-person appointments. The GPT-4 based model **analyses** data points from patient's diagnostic data to provide personalized solutions. By automating physicians' tasks, LLM enhance the platform's efficiency and accuracy, providing decision support for physicians. [88]

Be My Eyes Virtual Volunteer is a digital virtual assistant tool designed for individuals who are blind or have low vision. Powered by GPT-4, this application leverages advanced **image-to-text** functionality to assist users in their daily activities. By capturing images of their surroundings and posing questions, the Virtual Volunteer utilizes the integrated GPT-4 model to gain a deeper contextual understanding and effectively function as a conversational agent, simulating human interaction. [89]

Epic Systems' EHR software has a vast user base, with over 305 million patients' EHR worldwide [90]. Apotti, the Finnish electronic health and social data record, also employs Epic for its operations [90,91]. Epic utilizes GPT-3 and GPT-4 models for **drafting message responses** to patients and for **data analysis** purposes [90].

Dot Compliance provides electronic **quality management system** (eQMS) solutions for medical device and pharmaceutical organizations. Their latest eQMS product, QMS Xpress, incorporates a ChatGPT-powered platform that offers corrective and preventive suggestions for quality managers. It monitors and **analyses** all processes conducted in the eQMS, ensuring that no quality issues go unnoticed. This technology improves overall quality and efficiency while reducing costly regulatory penalties, recalls, and reputational damage. [92]

Kahun's **pre-consultation** tool integrates an evidence-based clinical reasoning tool with ChatGPT. The pre-consultation interview is utilized by using ChatGPT's engines, forming the basis for the subsequent clinical assessment. [93] Kahun also automatically integrates with the patient's EHR, connecting the original sources to the clinical reasoning path and recommending related workup for physicians. The tool offers trustworthy, patient-tailored clinical assessment to optimize physician time, along with a user-oriented interface made possible by ChatGPT integration. [94]

4.3 Performance of current models

This section aims to delve into the performance and potential of LLMs in healthcare applications, while also discussing their accuracy rates, effectiveness, and benchmark scores. The objective is to examine the variations and distinctions among different LLMs addressed in this study.

4.3.1 Diagnostics and triage

In a study of Levine et al., **GPT-3**'s **diagnostic** and **triage** performances were compared with those of lay Internet users and primary care physicians. Using 48 validated case vignettes for common and severe illnesses, the results indicated that GPT-3 performed significantly better than lay individuals but fell short of the performance exhibited by physicians. [95] These findings suggest that while GPT-3 is not comparable to physicians on its own, it may have potential in assisting lay individuals with health-related tasks.

Agbavor et al. utilized **GPT-3** to assist in the early **diagnosis** of dementia using speech data. The GPT-3 model generated text embeddings of transcribed speech, capturing the semantic meaning and providing a viable approach to predicting dementia from spontaneous speech. Their study demonstrated the potential utility of GPT-3 in detecting dementia and interpreting a patient's cognitive abilities [74].

4.3.2 Radiology

Wagner et al. conducted an evaluation of the performance of **ChatGPT-3** in **radiological QA**. In their study, ChatGPT-3 provided correct answers to only two-thirds of the questions. Additionally, they discovered that the answers provided by ChatGPT-3 often lacked proper references or contained insufficient information to address the questions adequately [96].

Rao and her colleagues evaluated **ChatGPT-3.5** in CDS for **radiology**. They evaluated the capability of ChatGPT-3.5 to identify appropriate imaging services for breast pain and breast cancer screening. They found that the accuracy varied by one-third, reaching approximately 90 % accuracy in breast cancer screening and under 60 % for breast pain. They also observed significant differences in accuracy based on the employed prompting strategies. [50] Similarly, Bhayana et al. explored **ChatGPT-3.5**'s performance in **radiology** using a dataset of 150 multiple-choice questions, resembling the

Canadian Royal College and American Board of Radiology examinations without images. They discovered that ChatGPT-3.5 performed better in questions requiring lower-order understanding compared to those requiring higher-order thinking, particularly struggling with questions involving description of imaging findings, calculation, classification, and concept application. Overall, it answered 69 % of questions correctly, nearly passing the radiology board-style examination without radiology-specific pretraining, suggesting exciting potential for LLMs in radiology. [65]

Jeblick and co-authors and Lyu et al. utilized **ChatGPT-3.5** to **simplify radiology reports**. Jeblick et al. simplified 45 reports, which were evaluated by 15 radiologists who generally deemed them factually correct, though some statements were incorrect or lacked key information [97]. In Lyu et al.'s study, radiologists evaluated 138 simplified screening reports generated by ChatGPT-3.5. The model received an average score of 4.268 on a five-point system, with a minimal incorrectness and missing information. However, the suggestions in the reports tended to be more general than specific. The study also investigated the impact of different prompts on performance, revealing opportunities for notable improvements. Additionally, they compared ChatGPT-3.5's performance with that of GPT-4 in the same task, and GPT-4 generated reports were evaluated as significantly higher quality, achieving 96.8 % accuracy with the optimized prompt. These results illustrate the potential of LLMs for report simplification but emphasize the need for physician involvement and fine-tuning of the used model. [66]

CheXbert is an LLM designed for automated **radiology report labelling**. It is fine-tuned from BERT using existing radiology report labellers and expert annotations augmented with backtranslation. CheXbert has demonstrated superior performance to previous models trained solely on radiologist labels or only on existing report labellers, nearly matching the performance of board-certified radiologists. [35] **CXR-BERT** is a radiology-specific text encoder trained from scratch, exhibiting improved performance in radiology NLI tasks, surpassing PubMedBERT, ClinicalBERT, and the score of the radiology NLI benchmark [27].

4.3.3 Benchmarks

Zhong and his team conducted a comprehensive comparison between **ChatGPT-3** and fine-tuned **BERT** models. Using the **GLUE** benchmark, they found that ChatGPT-3 performed better in inference tasks due to its reasoning ability, while BERT models outperformed ChatGPT-3 in **NLU** tasks such as paraphrase and similarity. In sentiment analysis and QA, both models performed equally. The study also highlighted the significant performance improvement when utilizing advanced prompting strategies with

ChatGPT-3. [98] Additionally, Tang et al. showed that **ChatGPT-3** underperformed compared to fine-tuned **BERT** models. They proposed a training paradigm to improve the performance of LLMs, but even with the improvement, it did not surpass the performance of BERT models. [40] Similarly, Gutiérrez et al. reported similar results, with the GPT-3 model failing to surpass fine-tuned BERT models [99].

Liévin et al. demonstrated that various LLMs, including GPT-3.5, U-PaLM, PubMedBERT, BioLinkBERT, BioGPT, BioMedLM (PubMedGPT), and Galatica achieved comparable performance to humans in MedQA, MedMCQA, and PubMedQA **benchmarks**. They also highlighted the significant improvement in model performance when using CoT in MQA. [46] In comparison, Au Yeung et al. showed that **ForesightGPT** outperformed **ChatGPT-3.5** in diagnosing form clinical histories, offering more transparent output and specific suggestions [45]. Additionally, Thirunavukarasu et al. found that the performance of ChatGPT-3.5 fell below the average passing mark on the Applied Knowledge Test of general practitioners in the UK, indicating the need for further development [100].

According to Wang et al. **GPT-4** outperformed Bard and GPT-3.5 in tasks such as **NLI**, **NER**, and **STS**. In **RE**, both GPT-4 and Bard performed comparably, surpassing the performance of GPT-3.5 depending on the dataset [1]. Additionally, GPT-4 demonstrated superior performance to GPT-3.5 in a multiple-choice MQA **benchmark**, including 30 % higher score on MedQA [54]. Wang et al. also introduced the **SQ** prompting technique, which outperformed standard and CoT prompting techniques, suggesting its utilization to maximize the effectiveness of LLMs in the healthcare domain [1].

Domain-specific LLMs, such as BioBERT, PubMedBERT, and BioGPT demonstrated superior performance compared to LLMs not trained with medical data, such as OpenAI's GPT models and Google AI's BERT models [54]. For instance, **BioBERT**, fine-tuned on biomedical text mining tasks like **NER**, **RE**, **MQA**, outperformed BERT in these tasks [17]. Similarly, **BioGPT** achieved performance comparable to BioBERT and other pre-trained biomedical LLMs, including PubMedBERT, BioLinkBERT, in three end-to-end RE **benchmarks**, and PubMedQA. [20]

GatorTron outperforms BioBERT, ClinicalBERT and BioMegaTron on three **benchmark** datasets. Yang et al. explored the impact of scaling up the number of parameters and training data size in GatorTron models. The largest GatorTron model, with 8.9 million parameters, achieved the best performance across **NER**, **RE**, **NLI**, and **MQA** tasks. The medium-sized model performed slightly better in **STS** tasks. Increasing the training data size also resulted in improved performance in most tasks, except for MQA. The main challenge for GatorTron is to identify key information from longer paragraphs. These findings suggest the applicability of GatorTron models in medical AI systems due to their observed performance. [12]

Flan-PaLM achieves 67.6 % accuracy on the MedQA **benchmark**, and it exhibits slightly superior performance in the MedMCQA dataset compared to Galatica, as well as comparable performance in PubMedQA alongside BioGPT and PubMedGPT. Furthermore, Flan-PaLM outperforms PaLM, Copher, Chinchilla, and Galatica, in the MMLU dataset. [2] However, it falls short compared to GPT-4 in multiple-choice MQA benchmark with four datasets [54]. Singhal et al. corroborate these finding on different model size variants, aligning with the conclusions drawn by Yang et al. Additionally, Signhal et al. introduce Med-PaLM, which incorporates instruction prompt tuning based on Flan-PaLM [2].

Med-PaLM emerges as the first model to surpass the passing score on the MedQA dataset. The enhanced version, **Med-PaLM-2**, achieves a 19 % higher score than the previous state-of-the-art (SOTA). Furthermore, Med-PaLM achieves SOTA results in several other MQA **benchmarks**. Singhal et al. assess the performance of Med-PaLM-2 on long-form questions and find that the model's answers are preferred over those provided by physicians in eight of nine axes, including factuality and reasoning capability. When comparing Med-PaLM-2 to Med-PaLM, it is evident that Med-PaLM-2 performs significantly better across various axes, including lower risk of harm. Moreover, when compared to other LLMs in multiple-choice benchmarks, Med-PaLM-2 achieves the best results. Singhal et al.'s comprehensive study provides compelling evidence of Med-PaLM-2's superior SOTA performance in both multiple-choice and long-form MQA tasks. [23]

Several studies utilized the MedQA benchmark dataset, obtained from USMLE, to assess the performance of LLMs. Figure 7 presents the accuracy of specific LLMs, expressed as percentages, within the MedQA benchmark. Additionally, the graph incorporates markers denoting the respective parameters of each model, providing a visual representation of their sizes.

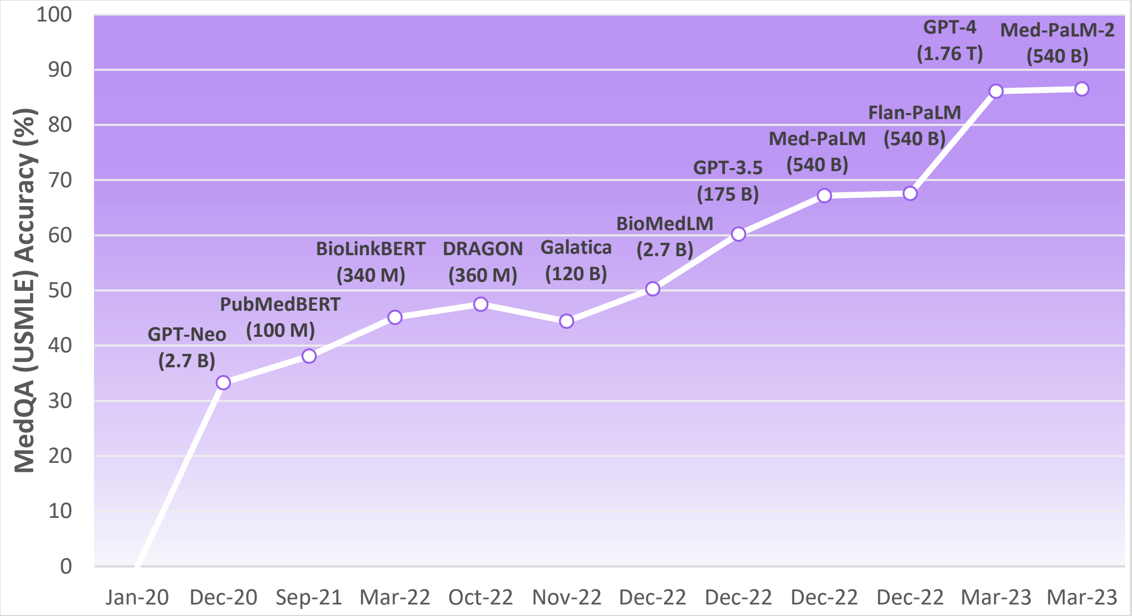


Figure 7: Accuracy (%) of various LLMs in the MedQA dataset containing multiple-choice questions from the US medical licencing exam. The respective numbers of parameters of each model are given in parentheses. [2,23]

The figure 7 reveals a clear upward trend over the past couple of years, indicating improvements in the model performance in USMLE accuracy. It can also be inferred that the size of the model is not directly proportional to its performance. For instance, BioMedLM outperformed the similarly sized GPT-Neo by 17 %. Additionally, Galatica performed approximately 3 % worse than DRAGON, despite being 300 times larger than DRAGON. Moreover, the two models with the highest accuracies, GPT-4 (86.1 %) and Med-PaLM-2 (86.5 %), differ in size, with GPT-4 being over 3 times larger than Med-PaLM-2.

4.3.4 Others

Chintagunta and co-workers employed **GPT-3** as the foundation of an algorithm for generating synthetic training data for **medical dialogue summarization** models. By combining medical knowledge with an ensemble of GPT-3 models, the algorithm generated synthetic labels that, when used alongside human-labelled data, produced high quality training data for summarization models. In comparison to models solely reliant on human-labelled data, the models trained using this proposed method exhibited enhanced precision and coherence in their summaries. Notably, the algorithm can effectively address privacy concerns by utilizing a limited and predetermined dataset. Given the algorithm's accuracy and privacy considerations, it holds practical potential for training medical dialogue summarization models. [61]

Ayers and colleagues conducted a study in which healthcare professionals compared the responses of physicians and **ChatGPT-3.5 chatbot** to patient's questions. The study found that the chatbot's responses were preferred over those of physicians. This is not surprising given the chatbot's ability to generate high-quality, empathetic text compared to busy physicians who aim to minimize the time spent on administrative work. [101]

Liu and her team employed **ChatGPT-3.5** to generate suggestions for improving **CDS alert** logic. CDS experts evaluated these suggestions, and the results indicated that ChatGPT-3.5 could complement traditional CDS optimization, analyse alert logic, and be integrated into the alert development stage. [55] Furthermore, Rao et al. compared standardized clinical vignettes for CDS and found that ChatGPT-3.5 performs better when provided with more clinical information [102].

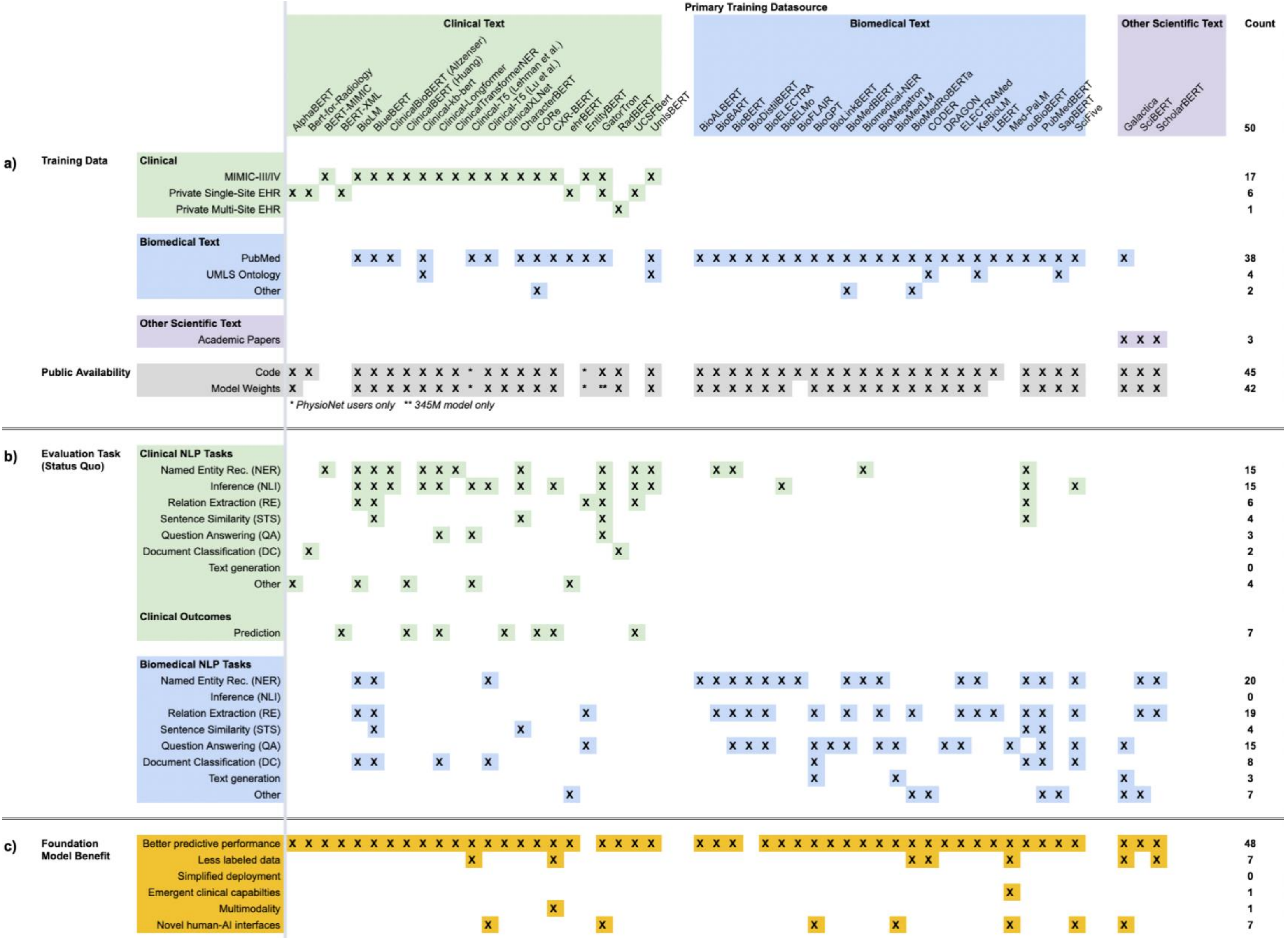
Yan and co-workers proposed an LLM based approach for **patient-trial matching**, employing **ChatGPT-4** to generate augmented data while preserving the semantic coherence of the original trial's eligibility criteria. They also utilized **BERT** as a text encoder for patient and eligibility criteria embeddings. The study demonstrated the effectiveness of LLMs adapted to patient-trial matching, resulting in an average improvement of 7.32 % in performance and 12.12 % in generalizability. [71]

InstructGPT was found to perform reasonably well in assisting physicians in determining patient eligibility for **clinical trials**. To further enhance its performance, Hamer et al. involved a physician to supervise the process, resulting in a 90 % reduction in workload during the pre-screening. [70]

Antikainen et al. evaluated the potential of **BERT** and **XLNet** in predicting mortality in cardiac patients using EHR data. Both models yielded similar results, with XLNet capturing more positive cases, but BERT achieving higher positive predictive value. Overall, both models achieved approximately 76 % accuracy, suggesting their potential integration into EHR systems in clinical practice. [103]

Lehman et al. conducted a study comparing the performance of **Clinical-T5-Base**, **Clinical-T5-Base-Cktp**, and **Clinical-T5-Large**, which are relatively small (ranging from 220 M to 770 M parameters), with larger models like GPT-3 (which has 175 B parameters). Their investigation revealed that these clinical domain models exhibit greater parameter efficiency compared to their larger counterparts. Furthermore, they observed that pretraining these models from scratch on clinical data positively impacts their overall performance. Additionally, all proposed domain-specific encoder-decoder models, including **SciFive** [42], demonstrate superior performance when compared to their general domain counterparts. However, it is worth noting that these encoder-decoder models are not compatible with other encoder-only or decoder-only models in biomedical **NLP tasks** [14]

Figure 8 illustrates various LLMs along with their corresponding training data and the public availability status of each model. The training data is categorized into clinical, biomedical, and other scientific texts. The MIMIC-III dataset comprises approximately two million medical notes from the ICU of the Beth Israel Deaconess Medical Centre between the years 2001 and 2012 [104]. Notably, training with MIMIC-III poses gaps in completeness due to the outdated information. Additionally, the figure presents the evaluation tasks and the benefits for which each model was assessed in its original research. This comprehensive figure effectively consolidates the discussed evaluation tasks and highlights the specific tasks for which each model can be utilized.



			Primary Training Datasource																						
			Clinical Text																						
			AlphaBERT	Bert-for-Radiology	BERT-MIMIC	BERT-XML	BioLM	BlueBERT	ClinicalBioBERT (Alzenser)	ClinicalBERT (Huang)	Clinical-kb-bert	Clinical-Longformer	ClinicalTransformerNER	Clinical-T5 (Lehman et al.)	Clinical-T5 (Lu et al.)	ClinicalXLNet	CharacterBERT	CORe	CXR-BERT	ehrBERT	EntityBERT	GatorTron	RadBERT	UCSFBert	UmlsBERT
a) Training Data	Clinical	MIMIC-III/IV			X		X	X	X	X	X	X	X	X	X	X	X	X	X		X	X			X
		Private Single-Site EHR	X	X		X														X		X		X	
		Private Multi-Site EHR																					X		
	Biomedical Text	PubMed					X	X	X		X			X	X		X	X	X	X	X	X			X
		UMLS Ontology									X														X
		Other																X							
	Other Scientific Text	Academic Papers																							
Public Availability		Code	X	X			X	X	X	X	X	X	X	*	X	X	X	X	X		*	X	X		X
		Model Weights	X				X	X	X	X	X	X	X	*	X	X	X	X	X		*	**	X		X
b) Evaluation Task (Status Quo)	Clinical NLP Tasks	Named Entity Rec. (NER)			X		X	X	X		X	X	X				X					X		X	X
		Inference (NLI)					X	X	X		X	X		X	X		X		X			X		X	X
		Relation Extraction (RE)					X	X													X	X		X	
		Sentence Similarity (STS)						X									X					X			
		Question Answering (QA)										X		X								X			
		Document Classification (DC)		X																			X		
		Text generation																							
		Other	X				X			X				X						X					
	Clinical Outcomes	Prediction				X				X		X				X		X	X					X	
	Biomedical NLP Tasks	Named Entity Rec. (NER)					X	X							X										
		Inference (NLI)																							
		Relation Extraction (RE)					X	X													X				
		Sentence Similarity (STS)						X									X								
		Question Answering (QA)																			X				
		Document Classification (DC)					X	X				X			X										
		Text generation																							
		Other																		X					
c) Foundation Model Benefit		Better predictive performance	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
		Less labeled data												X					X						
		Simplified deployment																							
		Emergent clinical capabilties																							
		Multimodality																	X						
		Novel human-AI interfaces													X							X			

			Primary Training Datasource																							
			Biomedical Text																							
			BioALBERT	BioBART	BioBERT	BioDistilBERT	BioELECTRA	BioELMo	BioFLAIR	BioGPT	BioLinkBERT	BioMedBERT	Biomedical-NER	BioMegatron	BioMedLM	BioMedRoBERTa	CODER	DRAGON	ELECTRAMed	KeBioLM	LBERT	Med-PaLM	ouBioBERT	PubMedBERT	SapBERT	SciFive
a) Training Data	Clinical	MIMIC-III/IV																								
		Private Single-Site EHR																								
		Private Multi-Site EHR																								
	Biomedical Text	PubMed	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		UMLS Ontology															X			X					X	
		Other										X				X										
	Other Scientific Text	Academic Papers																								
Public Availability		Code	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
		Model Weights	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X			X	X	X	X
b) Evaluation Task (Status Quo)	Clinical NLP Tasks	Named Entity Rec. (NER)		X	X								X										X			
		Inference (NLI)						X															X			X
		Relation Extraction (RE)																					X			
		Sentence Similarity (STS)																					X			
		Question Answering (QA)																								
		Document Classification (DC)																								
		Text generation																								
		Other																								
	Clinical Outcomes	Prediction																								
	Biomedical NLP Tasks	Named Entity Rec. (NER)	X	X	X	X	X	X	X			X	X	X					X	X			X	X		X
		Inference (NLI)																								
		Relation Extraction (RE)		X	X	X	X			X		X		X		X			X	X	X		X	X		X
		Sentence Similarity (STS)																					X	X		
		Question Answering (QA)			X	X	X			X	X	X		X	X			X	X			X		X		X
		Document Classification (DC)								X													X	X		X
		Text generation								X					X											
		Other														X	X							X	X	
c) Foundation Model Benefit		Better predictive performance	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
		Less labeled data														X	X					X				
		Simplified deployment																								
		Emergent clinical capabilties																				X				
		Multimodality																								
		Novel human-AI interfaces								X					X							X				X

			Primary Training Datasource			
			Other Scientific Text			Count
			Galactica	SciBERT	ScholarBERT	50
a) Training Data	Clinical	MIMIC-III/IV				17
		Private Single-Site EHR				6
		Private Multi-Site EHR				1
	Biomedical Text	PubMed	X			38
		UMLS Ontology				4
		Other				2
	Other Scientific Text	Academic Papers	X	X	X	3
Public Availability		Code	X	X	X	45
		Model Weights	X	X	X	42
b) Evaluation Task (Status Quo)	Clinical NLP Tasks	Named Entity Rec. (NER)				15
		Inference (NLI)				15
		Relation Extraction (RE)				6
		Sentence Similarity (STS)				4
		Question Answering (QA)				3
		Document Classification (DC)				2
		Text generation				0
		Other				4
	Clinical Outcomes	Prediction				7
	Biomedical NLP Tasks	Named Entity Rec. (NER)		X	X	20
		Inference (NLI)				0
		Relation Extraction (RE)		X	X	19
		Sentence Similarity (STS)				4
		Question Answering (QA)	X			15
		Document Classification (DC)				8
		Text generation	X			3
		Other	X	X		7
c) Foundation Model Benefit		Better predictive performance	X	X	X	48
		Less labeled data	X		X	7
		Simplified deployment				0
		Emergent clinical capabilties				1
		Multimodality				1
		Novel human-AI interfaces	X			7

* PhysioNet users only ** 345M model only

Figure 8: training, evaluation, and publication of various healthcare LLMs [105]

Upon examining Figure 8, it becomes evident that most of the models are trained solely on PubMed abstracts and/or full-text articles, either independently or in combination with other datasets. Among the models trained using clinical text, a significant portion utilized the MIMIC-III dataset. Additionally, it is noteworthy that nearly all models can be accessed publicly from online repositories such as HuggingFace. However, it is essential to highlight that while GatorTron has showcased exceptional performance, its accessibility to the public remains limited due to its training on private EMR datasets. Similarly, Med-Palm, another model exhibiting superior performance, has not disclosed the model's weights nor the underlying code for the public access. [105]

5. CHALLENGES OF LARGE LANGUAGE MODELS IN HEALTHCARE AND DISEASE DIAGNOSTICS

Despite the significant advancements and promising potential of LLMs in the healthcare field, the integration of these applications into actual medical practice remains limited. To responsibly and safely develop LLMs into the practical applications, it is essential to consider the ethical, technical, and cultural aspects [57]. The guiding principles for addressing ethical issues in the responsible integration of AI with medicine come from the World Health Organization [106]. This chapter will discuss the challenges and limitations of current LLMs when employing them in healthcare to supplement human expertise and how to mitigate them.

One major concern is the secondary use of medical data, such as EHRs or medical reports, as training data of LLMs or as input data in LLMs, which raises **privacy** concerns [54]. Medical data may not be publicly available due to regulations or the patient’s privacy concerns [61]. Patients are often uncertain about the secondary usage of their personal information, fearing inappropriate usage, stigmatization, or discrimination. Since medical data is highly sensitive and private, privacy concerns can affect the physician-patient relationship, inhibiting the sharing of necessary clinical information. [63] Even if the medical data is de-identified before the secondary usage, the risk of re-identifying patients from text data concerns both policymakers and the public, resulting in limited sharing of medical data outside of the clinical environment [107]. Additionally, utilizing medical papers in LLMs’ development without obtaining explicit consent raises privacy challenges [57].

The functions of LLMs are highly opaque, leading to **transparency** issues for both developers and users. The underlying logic between the prompt and output often remains obscure, consequently reducing the reliability of LLM’s deductions. [63] However, developers have the responsibility to promote transparency by openly sharing their methodologies, data sources, and potential biases. This practice is essential to enhance the understanding of the LLMs among users and regulators. [108]

Moreover, most LLMs lack proper **referencing** for their generated content. Even when references are provided, they may not accurately represent the output due to the influence of noisy training data [54]. For instance, Galatica was known to produce made-up references in its responses, leading to a rapid loss of trust in the model's credibility. Similarly, ChatGPT may generate references that only appear plausible but lack actual authenticity [51]. This poses a significant threat as spreading misinformation as truth can lead to permanent contamination of knowledge bases [57]. Additionally, LLMs can inadvertently engage in plagiarism as they might copy phrases from other documents without providing appropriate references, potentially leading to intellectual property and credibility issues [62].

LLMs, especially those trained for general domains, are meant to generate content that appears factual rather than producing strictly factual content [90]. The **accountability** of these models is limited to the used training data, as most models do not have access to the internet. Given that general domain LLMs are not specifically trained with clinical databases, questions may arise regarding the accuracy of medical facts. Additionally, the lack of common sense, variation of semantic expressions, and complexity of proper background information make LLMs prone to errors [63]. The strategies used by LLMs to improve the diversity of outputs can increase the occurrence of factual errors [54]. Furthermore, the lack of transparency regarding the training data used in some models, including ChatGPT's, makes it challenging to evaluate the originality of the data [96].

One major issue with LLMs is their tendency to amplify human **bias**. Biases can be reflected in LLMs directly from the training data collection and preparation. Biases can also emerge from the algorithm's design if it gives higher priority to certain data points. Unexpected biases may arise from the model's architecture, resulting from the interaction between parameters and biased training data. Moreover, if the training process involves human feedback, the subjective viewpoints of individuals providing feedback can unintentionally influence the model's behaviour, potentially leading to biased outcomes. Furthermore, the biases of LLMs are exacerbated by the policies set forth by their developers, affecting the values and decisions made by the models. [72,108]

These biases within LLMs are particularly concerning, especially in the healthcare field, as they can lead to inaccurate predictions, erroneous recommendations, misdiagnosis, and unequal access to care. Additionally, biases can amplify stereotypes, exacerbate disparities, and perpetuate inequalities within subpopulations, favouring certain individuals based on attributes such as gender, race, ethnicity, ideology, politics or other factors. [72,108]

LLMs tend to be more proficient in English due to the abundance of English content on the internet, leading to linguistic biases. In clinical settings, local languages are often used, demanding equal performance of LLMs in all languages. Additionally, the models may lack understanding of historical or current issues due to training on specific data periods. Biases embedded in base models, such as BERT or GPT are carried over to variant models based on them, which expands the issue [109].

Despite the improved performance of fine-tuned large architecture models such as GatorTron or Med-PaLM, they remain vulnerable to biases [110]. Collaboration between developers, clinicians, and users is required to **mitigate** and **address biases** [108]. Some models, like ChatGPT, prioritize data accessibility and availability over bias mitigation. Eliminating bias from models is complicated since they have already learned from biased data. Even when a model is tested for bias, such as Med-PaLM, constant auditing is required due to its probabilistic nature. [57]

To use LLMs as reliable and accountable tools, end-users, such as clinicians or patients, must be **educated** about the capabilities, limitations, and risks of LLMs. Users should not rely on them as omniscient tools and must be critical in their interactions. [57] Proper training of users is essential to prompt the model accurately and avoid **hallucination**, where the model generates factually incorrect output [45]. Generated content should be marked as AI-generated to avoid misinterpretation in clinical decision-making. Another potential concern related to hallucination is that the developers of LLMs may not assume full responsibility for the generated outputs, which can lead to uncertainties regarding accountability for poor outcomes. [57] Support for users on managing hallucinations and biases is crucial, and collaboration between manufacturers and users is key to success [108].

Regulations play a vital role in ensuring the safe, ethical, and effective development of these new AI-based tools. Due to their adaptive nature, the algorithms change continuously. In addition to the dynamic behaviour, the models' scale and complexity sets them apart from any previous deep learning methods. The U.S. Food and Drug Administration's FDA have started regulating software as a medical device, specifically to address AI and ML technologies throughout their lifespan. FDA has not yet solved the regulation for algorithms, that are adaptive and utilizes self-supervised learning, such as LLMs. [76] The American Medical Association addresses appropriate integration of AI into healthcare through the Augmented Intelligence in Health Care policy [63]. The European Union's General Data Protection Regulation GDPR applies to LLMs that process personal data within or from EU, ensuring privacy protection [111]. Findata, an act on the secondary use of health and social data, aims for secure processing of individuals’ personal data in Finland [112]. In Europe, LLMs intended for medical purposes may be classified as medical devices, and must adhere to the European Union Medical Device Regulation MDR, indicated by the CE marking [113]. Additionally, the Health Insurance Portability and Accountability Act HIPAA aims to protect individuals’ privacy and security in healthcare [114]. Regular audits of data privacy and security policies and regulations are necessary to ensure the responsible utilization of LLMs [57].

The categorization and regulatory oversight of LLMs intended for medical purposes pose intricate challenges, necessitating the establishment of a distinct regulatory category. Continuous monitoring and updates are imperative to keep pace with the rapid advancement in these models. Furthermore, integrating LLMs into certified medical technologies raises questions about their regulatory control. Despite the encountered difficulties, effective implementation of regulations can significantly alleviate safety, privacy, bias, and ethical concerns. [76]

Developing high-quality LLM systems in healthcare applications requires employing numerous highly skilled clinical experts to ensure model accuracy. However, this results in remarkable labour **costs**, which are often underestimated, leading to poor working conditions. [57] The cost of LLMs depends on the training procedure. While training from scratch incurs a high one-time cost, it reduces the cost of fine-tuning and running inference. On the other hand, models trained using general domain models as a foundation do not require pretraining costs, but may lead to more expensive inference and fine-tuning costs due to their larger size and technical expertise and infrastructure requirements. [14] Additionally, training large architecture models consumes large compute-power resources, contributing to a large carbon footprint, making sustainability an important consideration in the training process [57].

Reinforcement Learning with Human Feedback (RLHF) has improved the safety and faithfulness of LLMs, but it raises cost concerns. Further research is needed to develop RLHF as a more cost-effective and resource-efficient method. An automatic method to evaluate the factual correctness of LLMs effectively would address this issue. [54]

Another noteworthy challenge in the integration of LLMs into healthcare practice is the inherent **conservatism** observed within the healthcare system in general. Legacy systems continue to be extensively utilized due to healthcare practitioners’ inherent wariness towards embracing novel technologies and demands regarding backwards compatibility with earlier investments. Experienced clinicians may exhibit resistance in deviating from established practices, preferring to adhere to traditional approaches rather than adopting the next generation of technology. Moreover, the challenges currently faced by LLMs act as deterrents for healthcare practitioners. Additionally, there are apprehensions regarding the potential unforeseen complications arising from the interaction between AI systems and humans [115].

To serve as reliable and accountable tools, LLM systems should be developed transparently, defining values and purposes, and ensuring adherence to the defined framework. Proper curation of training data, representing diverse insights, backgrounds, and histories, is crucial to minimize biases. Developing LLMs is a continuous process that involves active participation from developers, regulators, users, and other stakeholders to constantly evaluate, refine, and improve the models. [57,108]

6. DISCUSSION

Originally, this study aimed to evaluate the performance of current LLMs in medical practice. However, due to LLMs being a relatively new technology and rarely used in practice yet, most of the current literature focuses on evaluating these models against standardized benchmark datasets to assess their performance in clinical language understanding tasks. Consequently, only a few studies have explored the practical usage of LLMs. Nevertheless, this literature review effectively demonstrates the current potential for LLM utilization in healthcare and comprehensively examines the prevailing challenges and potential solutions.

The assessment of the value of LLMs presents a complex undertaking. On one hand, general domain language models face limitations in effectively capturing medical knowledge due to the scarcity of biomedical data in their training. Conversely, domain-specific LLMs suffer from the restricted diversity of available biomedical training data, primarily because biomedical texts are not publicly accessible due to privacy concerns. Consequently, domain-specific models, solely trained on biomedical data, and fine-tuned general models both have limitations. [54]

Notwithstanding the challenges in data availability for domain-specific models, they are considered more parameter-efficient compared to general models. Even the integration of ICL fails to elevate the performance of general models to the level of specialized models. [39] Parameter-efficient models not only mitigate training expenses but also expedite the training process. Thus, further development of highly specialized models is imperative to cater to the precise demands of healthcare. Addressing privacy concerns and elevating LLMs' performance to new heights necessitates the establishment of proper regulations.

The development of these specialized models requires a profound understanding of model architectures. Traditionally, the choice of architecture has been driven by the target task. For instance, encoder-only models have been employed for tasks requiring text comprehension or information extraction, while decoder-only models have been utilized in text generation, translation, and QA. Encoder-decoder models have found application in tasks where capturing relationships between the input and output is critical, such as text summarization. This study showed that although encoder-decoder models are designed for these tasks, they have not been widely developed or used.

Future research could investigate why these models have lagged encoder-only and decoder-only models. Furthermore, this study has indicated that these models can exhibit impressive performance not only in their designated tasks but also in other tasks, as evidenced by GPT models excelling in text classification, sentiment analysis, and summarization.

The prompting strategies should also be employed in the future to guide and refine the behaviour for optimized performance of LLMs. It has been demonstrated that employing prompting strategies considerable enhances model performance, yet the strategy and the prompting quality can affect the effectiveness. For example, the SQP introduced by Wang et al. and the ER proposed by Singhal et al. demonstrated improved performance compared to standard and CoT strategies in healthcare settings [1,23]. Prospective research could potentially contribute to the advancement of novel and more efficient prompting strategies.

Despite GPT models being widely acclaimed for their adaptability in various NLU tasks, encoder models, such as fine-tuned BERT variants, exhibit significant value in specific NLU tasks like RE and NER, where they outperform decoder models. Both GPT models and fine-tuned BERT did show promising potential in radiology applications. Notably, the study conducted by Lyu et al. highlighted the eminent performance of GPT-4 with an optimized prompt, achieving an impressive 96.8 % accuracy, despite not being specifically trained for radiology [66]. In addition to the remarkable performance of GPT-4 and Med-PaLM-2, GatorTron has exhibited notable potential as well. The advancements in LLMs are moving in the right direction; however, current research highlights a lack of proper referencing, which poses challenges to the transparent supervision of these models. Addressing these concerns would further enhance the utility of current models in practical applications.

Currently, LLMs confront numerous challenges, and it is noteworthy that there are no specific regulatory guidelines addressing them. Nevertheless, millions of individuals, including doctors and patients rely on LLMs daily. Moreover, the tokenization process, which plays a key role in how LLMs handle NLP, remains unregulated in the healthcare domain. [116] Apart from privacy protection, regulations could also help address various ethical concerns. Therefore, regulators should create a new regulatory category tailored to LLMs, given their distinctive nature from other medical technologies. Wise development of regulations is essential to accommodate the continuous evolution of these extensive and dynamic models, which have the potential to revolutionize our society. Regulators should also offer guidance to healthcare organizations, institutions, and companies on responsible LLM deployment. [76]

In fact, the active involvement of all stakeholders is necessary for the development of LLMs that effectively serve the demanding healthcare domain. Researchers should devise solutions to overcome the current challenges, including accuracy improvements, reduction of hallucinations, addressing bias and ethical considerations, and ensuring patient data privacy and safety. For instance, further research could be dedicated to developing more effective debiasing methods, which have not yet succeeded in mitigating biases [36]. Companies should boldly experiment with LLMs and innovate to integrate them into their products. Trial and error are crucial for the successful adoption of LLMs in healthcare applications. However, regulations should guide this process to ensure reliable LLM operation and to maintain human control throughout the application's lifecycle. Clinicians and patients should collaborate with researchers and developers in the model development process, as they are the primary end-users. Additionally, in the implementation of the LLM applications, the end-users should be educated for the proper, effective, and safe usage.

Most LLMs are trained using datasets limited to specific time periods. However, new diseases, treatments, and practices are continually being discovered, making LLMs employed in healthcare vulnerable to performance gaps if their training data becomes outdated. Therefore, clinical LLMs should undergo constant updates with current medical data, and applications utilizing these models should always be kept up-to-date to match the latest information. Regulations could also govern this perspective by restricting the use of LLMs trained on outdated data in healthcare applications.

The current healthcare system faces diverse challenges, including financial constraints, managing increasing volumes of medical data, accessibility and equity concerns, and a raising need for healthcare services. LLMs have the potential to mitigate these challenges by optimizing healthcare operations, automating administrative tasks to reduce costs, processing, and analysing healthcare data rapidly to increase efficiency, and allowing healthcare professionals to focus on their core tasks. LLMs can also improve access to healthcare by enabling remote medical care and enhancing patient-clinician engagement. Moreover, education capabilities and the CDS can elevate the quality of care. Applications likely to be employed for documentation purposes are anticipated to be the first to see practical implementation due to their ability to effectively address current healthcare challenges. Their associated risks are also minimal, as they do not generate novel information.

Physicians possess significant leverage to advocate for and influence political, economic, and social decisions. However, due to limited time and lack of advocacy training, physicians often fail to utilize this influence effectively. LLMs could play a role in this aspect by drafting tailored and appropriate communications to lawmakers. [51]

Apart from improving LLMs' performance in their current tasks, they are likely to acquire entirely new capabilities. For example, the latest release of OpenAI, GPT-4, can process images alongside text, potentially offering unforeseen solutions for healthcare through image analysis. In the future, LLMs may be able to analyse sound and video formats, further expanding their versatile and comprehensive capabilities. Moreover, novel and innovative solutions using LLMs in less automation-prone healthcare applications are expected to emerge [116].

7. CONCLUSIONS

The objective of this study was to explore the potential and challenges of LLMs in healthcare and CDS applications. It was evident from the findings that developing healthcare domain-specific LLMs is imperative to meet the unique demands of the field. The performance evaluation demonstrated the substantial capacity of LLMs to enhance the effectiveness and accessibility of healthcare. Furthermore, the model architecture and prompting strategies were identified as critical factors impacting the suitability of LLMs. These aspects should be carefully considered during the development of healthcare LLMs to ensure their reliability and accuracy.

Notwithstanding the significant progress made in LLMs, their widespread adoption into medical practice faces impediments posed by a range of challenges. These challenges encompass privacy concerns, transparency issues, accountability problems, potential hallucinations, and bias amplification. Moreover, the absence of tailored regulations for LLMs contributes remarkably to these issues. The implementation of specific regulatory guidelines tailored to LLMs would certainly ameliorate the current challenges, enhancing the accountability, ethics, and safety of LLMs in healthcare applications.

Although LLMs are currently scarcely used in medical practice, this study has demonstrated their promising potential to revolutionize the field in the future. While notable obstacles remain, the ongoing development of LLMs shows promise, and the current hype surrounding them suggests continuous improvement. To address all aspects comprehensively, the involvement of various stakeholders, including healthcare professionals, regulators, institutions, companies, and patients, is essential during the development process.

LLMs have the capacity to significantly enhance the efficiency of healthcare professionals by automating administrative tasks and analysing healthcare data. Additionally, they hold promise in optimizing CDS systems, thereby increasing the accuracy and speed of decision-making processes. For patients, LLM-powers chatbots can serve as virtual assistants, providing health guidance and enhancing healthcare accessibility. Embracing LLMs in these applications would be invaluable in managing vast volumes of medical data and meeting the growing demand for healthcare services.

This study has successfully illustrated the current state of healthcare LLMs and their potential for future development to address current challenges and facilitate their utilization in medical practice. Future research efforts could focus on advancing LLMs to improve performance while considering the factors identified in this study to influence their effectiveness. Additionally, solutions should be sought to address the challenges unveiled in this study. Given the rapid evolution of LLMs and the challenges identified, the subsequent years will determine whether their utilization in healthcare remains merely a subject of hype or ultimately leads to a revolutionary transformation of healthcare technology.

REFERENCES

[1] Wang Y, Zhao Y, Petzold L. Are Large Language Models Ready for Healthcare? A Comparative Study on Clinical Language Understanding. ArXivOrg 2023.

[2] Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large Language Models Encode Clinical Knowledge 2022. https://doi.org/10.48550/arXiv.2212.13138.

[3] Pathak P. Large Language Models 101: History, Evolution and Future. Scribble Data 2023. https://www.scribbledata.io/large-language-models-history-evolutions-and-future/ (accessed July 23, 2023).

[4] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need 2017.

[5] ChatGPT Statistics 2023 Revealed: Insights & Trends 2023. https://blog.gitnux.com/chat-gpt-statistics/ (accessed July 23, 2023).

[6] PRISMA n.d. http://www.prisma-statement.org/PRISMAStatement/FlowDiagram?AspxAutoDetectCookieSupport=1 (accessed July 26, 2023).

[7] Zotero | Your personal research assistant n.d. https://www.zotero.org/ (accessed July 27, 2023).

[8] Chen M, Decary M. Artificial intelligence in healthcare: An essential guide for health leaders. Healthc Manage Forum 2020;33:10–8. https://doi.org/10.1177/0840470419873123.

[9] The rise of artificial intelligence in healthcare applications - Tampere University Foundation n.d. https://andor.tuni.fi/discovery/fulldisplay/cdi_scopus_primary_637306417/358FIN_TAMPO:VU1 (accessed May 25, 2023).

[10] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24–9. https://doi.org/10.1038/s41591-018-0316-z.

[11] Ongsulee P. Artificial intelligence, machine learning and deep learning. 2017 15th Int. Conf. ICT Knowl. Eng. ICTKE, 2017, p. 1–6. https://doi.org/10.1109/ICTKE.2017.8259629.

[12] Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. Npj Digit Med 2022;5:1–9. https://doi.org/10.1038/s41746-022-00742-2.

[13] Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. Nat Biomed Eng 2022;6:1346–52. https://doi.org/10.1038/s41551-022-00914-1.

[14] Do We Still Need Clinical Language Models? - ProQuest n.d. https://www.proquest.com/docview/2777527015?pq-origsite=primo&accountid=14242 (accessed May 22, 2023).

[15] PhD PP. A fascinating tree of GPTs and LLMs reveals what's been going on. Medium 2023. https://medium.com/@paul.k.pallaghy/a-fascinating-tree-of-gpts-and-llms-reveals-whats-been-going-on-4d4235f2a2b1 (accessed June 14, 2023).

[16] OpenAI Platform n.d. https://platform.openai.com (accessed July 14, 2023).

[17] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–40. https://doi.org/10.1093/bioinformatics/btz682.

[18] Hagen A. Domain-specific language model pretraining for biomedical natural language processing. Microsoft Res 2020. https://www.microsoft.com/en-us/research/blog/domain-specific-language-model-pretraining-for-biomedical-natural-language-processing/ (accessed July 14, 2023).

[19] Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings 2019.

[20] Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23. https://doi.org/10.1093/bib/bbac409.

[21] Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: Scaling Language Modeling with Pathways 2022.

[22] Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling Instruction-Finetuned Language Models 2022.

[23] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al. Towards Expert-Level Medical Question Answering with Large Language Models 2023.

[24] Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance 2022. https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html (accessed July 14, 2023).

[25] Shin H-C, Zhang Y, Bakhturina E, Puri R, Patwary M, Shoeybi M, et al. BioMegatron: Larger Biomedical Domain Language Model. Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. EMNLP, Online: Association for Computational Linguistics; 2020, p. 4700–6. https://doi.org/10.18653/v1/2020.emnlp-main.379.

[26] Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining Language Models with Document Links 2022.

[27] Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. Comput. Vis. - ECCV 2022, vol. 13696, Switzerland: Springer; 2022, p. 1–21. https://doi.org/10.1007/978-3-031-20059-5_1.

[28] BioMedLM: a Domain-Specific Large Language Model for Biomedical Text n.d. https://www.mosaicml.com/blog/introducing-pubmed-gpt (accessed July 14, 2023).

[29] Yasunaga M, Bosselut A, Ren H, Zhang X, Manning CD, Liang P, et al. Deep Bidirectional Language-Knowledge Graph Pretraining 2022.

[30] Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text 2019.

[31] Papanikolaou Y, Pierleoni A. DARE: Data Augmented Relation Extraction with GPT-2 2020.

[32] Taylor R, Kardas M, Cucurull G, Scialom T, Hartshorn A, Saravia E, et al. Galactica: A Large Language Model for Science 2022.

[33] Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Au J, et al. Foresight - Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs n.d.

[34] Foresight n.d. https://foresight.sites.er.kcl.ac.uk/ (accessed July 14, 2023).

[35] Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT 2020.

[36] Lalor J, Yang Y, Smith K, Forsgren N, Abbasi A. Benchmarking Intersectional Biases in NLP. Proc. 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol., Seattle, United States: Association for Computational Linguistics; 2022, p. 3598–609. https://doi.org/10.18653/v1/2022.naacl-main.263.

[37] Chinchilla AI by Deepmind Review 2023. Writecream 2023. https://www.writecream.com/chinchilla-review-2023/ (accessed July 14, 2023).

[38] An empirical analysis of compute-optimal large language model training n.d. https://www.deepmind.com/publications/an-empirical-analysis-of-compute-optimal-large-language-model-training (accessed July 14, 2023).

[39] FERMA | The quickest path to your next eureka. FERMA n.d. https://ferma.ai/ (accessed July 4, 2023).

[40] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer 2020.

[41] Lewis P, Ott M, Du J, Stoyanov V. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. Proc. 3rd Clin. Nat. Lang. Process. Workshop, Online: Association for Computational Linguistics; 2020, p. 146–57. https://doi.org/10.18653/v1/2020.clinicalnlp-1.17.

[42] Phan LN, Anibal JT, Tran H, Chanana S, Bahadroglu E, Peltekian A, et al. SciFive: a text-to-text transformer model for biomedical literature. ArXivOrg 2021.

[43] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding 2020.

[44] Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, et al. Scaling Language Models: Methods, Analysis & Insights from Training Gopher 2022.

[45] Au Yeung J, Kraljevic Z, Luintel A, Balston A, Idowu E, Dobson RJ, et al. AI chatbots not yet ready for clinical use. Front Digit Health 2023;5:1161098. https://doi.org/10.3389/fdgth.2023.1161098.

[46] Liévin V, Christoffer Egeberg Hother, Winther O. Can large language models reason about medical questions? ArXivOrg 2023.

[47] EMR vs. EHR: Understand the Difference | Elevance Health. WwwElevancehealthCom n.d. https://www.elevancehealth.com/our-approach-to-health/digitally-enabled-healthcare/know-the-difference-between-ehr-and-emr (accessed July 3, 2023).

[48] Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? Am J Med 2018;131:129–33. https://doi.org/10.1016/j.amjmed.2017.10.035.

[49] Marr B. Revolutionizing Healthcare: The Top 14 Uses Of ChatGPT In Medicine And Wellness. Forbes n.d. https://www.forbes.com/sites/bernardmarr/2023/03/02/revolutionizing-healthcare-the-top-14-uses-of-chatgpt-in-medicine-and-wellness/ (accessed June 29, 2023).

[50] Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. MedRxiv Prepr Serv Health Sci 2023:2023.02.02.23285399. https://doi.org/10.1101/2023.02.02.23285399.

[51] Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology 2023;307:e230163. https://doi.org/10.1148/radiol.230163.

[52] About us. SNOMED Int n.d. https://www.snomed.org/about-us (accessed July 5, 2023).

[53] Moons P, Van Bulck L. ChatGPT: can artificial intelligence language models be of value for cardiovascular nurses and allied health professionals. Eur J Cardiovasc Nurs 2023:zvad022. https://doi.org/10.1093/eurjcn/zvad022.

[54] Xie Q, Wang F. Faithful AI in Healthcare and Medicine. MedRxiv 2023. https://doi.org/10.1101/2023.04.18.23288752.

[55] Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Assessing the Value of ChatGPT for Clinical Decision Support Optimization. MedRxiv Prepr Serv Health Sci 2023:2023.02.21.23286254. https://doi.org/10.1101/2023.02.21.23286254.

[56] Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models. ArXivOrg 2023.

[57] 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. https://doi.org/10.1016/j.ebiom.2023.104512.

[58] Clinical Notes n.d. http://www.healthit.gov/isa/uscdi-data-class/clinical-notes (accessed June 30, 2023).

[59] Biswas S. ChatGPT and the Future of Medical Writing. Radiology 2023;307:e223312. https://doi.org/10.1148/radiol.223312.

[60] Canes D. The Time-Saving Magic of Chat GPT for Doctors. Till Cavalry Arrive 2022. https://tillthecavalryarrive.substack.com/p/the-time-saving-magic-of-chat-gpt (accessed May 25, 2023).

[61] Chintagunta B, Katariya N, Amatriain X, Kannan A. Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization. ArXivOrg 2021.

[62] Kitamura FC. ChatGPT Is Shaping the Future of Medical Writing But Still Requires Human Judgment. Radiology 2023;307:e230171. https://doi.org/10.1148/radiol.230171.

[63] Xu L, Sanders L, Li K, Chow JCL. Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. JMIR Cancer 2021;7:e27850. https://doi.org/10.2196/27850.

[64] Bajaj RHD Simar S. Promises — and pitfalls — of ChatGPT-assisted medicine. STAT 2023. https://www.statnews.com/2023/02/01/promises-pitfalls-chatgpt-assisted-medicine/ (accessed May 24, 2023).

[65] Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology 2023:230582. https://doi.org/10.1148/radiol.230582.

[66] Lyu Q, Tan J, Zapadka ME, Ponnatapura J, Niu C, Myers KJ, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6:9. https://doi.org/10.1186/s42492-023-00136-5.

[67] Clinical Practice Guidelines : Writing a good medical report n.d. https://www.rch.org.au/clinicalguide/guideline_index/Writing_a_good_medical_report/ (accessed July 3, 2023).

[68] Clifford Stermer, MD (@tiktokrheumdok). TikTok n.d. https://www.tiktok.com/@tiktokrheumdok (accessed July 13, 2023).

[69] Stuart Blitz [@StuartBlitz]. You: There's no ChatGPT use case in healthcare Docs: Watch this 👇 https://t.co/2hX6vT4ncn. Twitter 2022.

[70] den Hamer DM, Schoor P, Polak TB, Kapitan D. Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models. ArXivOrg 2023.

[71] Yuan J, Tang R, Jiang X, Hu X. LLM for Patient-Trial Matching: Privacy-Aware Data Augmentation Towards Better Performance and Generalizability. ArXivOrg 2023.

[72] Pal R, Garg H, Patel S, Sethi T. Bias Amplification in Intersectional Subpopulations for Clinical Phenotyping by Large Language Models 2023:2023.03.22.23287585. https://doi.org/10.1101/2023.03.22.23287585.

[73] Amini S, Hao B, Zhang L, Song M, Gupta A, Karjadi C, et al. Automated detection of mild cognitive impairment and dementia from voice recordings: A natural language processing approach. Alzheimers Dement 2022;19:946–55. https://doi.org/10.1002/alz.12721.

[74] Agbavor F, Liang H. Predicting dementia from spontaneous speech using large language models. PLOS Digit Health 2022;1:e0000168. https://doi.org/10.1371/journal.pdig.0000168.

[75] Willett FR, Avansino DT, Hochberg LR, Henderson JM, Shenoy KV. High-performance brain-to-text communication via handwriting. Nature 2021;593:249–54. https://doi.org/10.1038/s41586-021-03506-2.

[76] Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. Npj Digit Med 2023;6:1–6. https://doi.org/10.1038/s41746-023-00873-0.

[77] Lunden I. Nabla, a digital health startup, launches Copilot, using GPT-3 to turn patient conversations into action. TechCrunch 2023. https://techcrunch.com/2023/03/14/nabla-a-french-digital-health-startup-launches-copilot-using-gpt-3-to-turn-patient-conversations-into-actionable-items/ (accessed June 26, 2023).

[78] Nabla Copilot · Superpowers for clinicians n.d. https://www.nabla.com/ (accessed June 26, 2023).

[79] Center MN. Microsoft accelerates industry cloud strategy for healthcare with the acquisition of Nuance. Stories 2021. https://news.microsoft.com/2021/04/12/microsoft-accelerates-industry-cloud-strategy-for-healthcare-with-the-acquisition-of-nuance/ (accessed June 27, 2023).

[80] Inc NC. Nuance and Microsoft Announce the First Fully AI-Automated Clinical Documentation Application for Healthcare n.d. https://www.prnewswire.com/news-releases/nuance-and-microsoft-announce-the-first-fully-ai-automated-clinical-documentation-application-for-healthcare-301775640.html (accessed June 27, 2023).

[81] LandiFeb 10 H, 2023 07:00pm. Doximity rolls out beta version of ChatGPT tool for docs aiming to streamline administrative paperwork. Fierce Healthc 2023. https://www.fiercehealthcare.com/health-tech/doximity-rolls-out-beta-version-chatgpt-tool-docs-aiming-streamline-administrative (accessed July 4, 2023).

[82] AI-powered, Physician-led Virtual Healthcare. Babylon Health n.d. https://www.babylonhealth.com/en-us/ (accessed July 26, 2023).

[83] Dare A. ChatGPT Healthcare: for Patients and Healthcare Professionals. MedMatch Netw 2023. https://medmatchnetwork.com/transforming-healthcare-access-how-chatgpt-and-medmatch-network-revolutionize-patient-care-chatgpt-healthcare/ (accessed June 26, 2023).

[84] Find Support Anytime, Anywhere with Our AI-Powered Mental Health ChatBot n.d. https://www.chatbeacon.io/industry-chatgpt/mental-health-chatbot (accessed July 3, 2023).

[85] (14) 16 Healthcare Companies That Already Integrated ChatGPT: Infographic | LinkedIn n.d. https://www.linkedin.com/pulse/16-healthcare-companies-already-integrated-chatgpt-mesk%25C3%25B3-md-phd/?trackingId=RocXXs0GQhSbkV30WN12qw%3D%3D (accessed May 19, 2023).

[86] Livewello | Genome Data Analysis n.d. https://livewello.com/ (accessed July 4, 2023).

[87] Amazfit Cheetah Round – amazfit-global-store n.d. https://www.amazfit.com/products/amazfit-cheetah-round (accessed July 4, 2023).

[88] Lunden I. Bionic Health raises $3M for its AI health clinic using GPT-4 and other ML models to design better preventative care. TechCrunch 2023. https://techcrunch.com/2023/03/21/bionic-health-raises-3m-for-its-ai-health-clinic-using-gpt-4-and-other-ml-models-to-design-better-preventative-care/ (accessed July 4, 2023).

[89] Introducing Our Virtual Volunteer Tool for People who are Blind or Have Low Vision, Powered by OpenAI's GPT-4 n.d. https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer (accessed July 3, 2023).

[90] Edwards B. GPT-4 will hunt for trends in medical records thanks to Microsoft and Epic. Ars Tech 2023. https://arstechnica.com/information-technology/2023/04/gpt-4-will-hunt-for-trends-in-medical-records-thanks-to-microsoft-and-epic/ (accessed June 27, 2023).

[91] Advancing Health and Social Care in Finland with Epic n.d. https://www.epic.com/epic/post/advancing-health-social-care-finland-epic (accessed June 27, 2023).

[92] Compliance D. Dot Compliance Launches First AI-Based ChatGPT Powered eQMS For Life Sciences n.d. https://www.prnewswire.com/news-releases/dot-compliance-launches-first-ai-based-chatgpt-powered-eqms-for-life-sciences-301789665.html (accessed June 27, 2023).

[93] Renolayan J. Kahun integrates ChatGPT, bolstering its AI that masters the fundamentals of medicine. Tech Times 2023. https://www.techtimes.com/articles/289851/20230402/pr-kahun-integrates-chatgpt-bolstering-ai-masters-fundamentals-medicine.htm (accessed June 27, 2023).

[94] Kahun.com n.d. https://www.kahun.com/ (accessed June 27, 2023).

[95] Dm L, R T, B K, A V, Sg F, A M, et al. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model. Medrxiv Prepr Serv Health Sci 2023:2023.01.30.23285067-2023.01.30.23285067. https://doi.org/10.1101/2023.01.30.23285067.

[96] Wagner MW, Ertl-Wagner BB. Accuracy of Information and References Using ChatGPT-3 for Retrieval of Clinical Radiological Information. Can Assoc Radiol J 2023:8465371231171125–8465371231171125. https://doi.org/10.1177/08465371231171125.

[97] Jeblick K, Schachtner B, Dexl J, Mittermeier A, Stüber AT, Topalis J, et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports 2022.

[98] Zhong Q, Ding L, Liu J, Du B, Tao D. Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT. ArXivOrg 2023.

[99] Gutiérrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, et al. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again 2022. https://doi.org/10.48550/arXiv.2203.08410.

[100] Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. 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. https://doi.org/10.2196/46599.

[101] Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Intern Med 2023. https://doi.org/10.1001/jamainternmed.2023.1838.

[102] Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow. MedRxiv Prepr Serv Health Sci 2023:2023.02.21.23285886. https://doi.org/10.1101/2023.02.21.23285886.

[103] Antikainen E, Linnosmaa J, Umer A, Oksala N, Eskola M, van Gils M, et al. Transformers for cardiac patient mortality risk prediction from heterogeneous electronic health records 2023. https://doi.org/10.1038/s41598-023-30657-1.

[104] Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.

[105] Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The Shaky Foundations of Clinical Foundation Models: A Survey of Large Language Models and Foundation Models for EMRs. ArXivOrg 2023.

[106] Ethics and governance of artificial intelligence for health n.d. https://www.who.int/publications-detail-redirect/9789240029200 (accessed July 18, 2023).

[107] Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. J Med Ethics 2020;46:367–77. https://doi.org/10.1136/medethics-2019-105472.

[108] Ferrara E. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. ArXivOrg 2023.

[109] Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings 2020. https://doi.org/10.48550/arXiv.2003.11515.

[110] Zakka C, Chaurasia A, Shad R, Hiesinger W. Almanac: Knowledge-Grounded Language Models for Clinical Medicine. ArXivOrg 2023.

[111] (22) The EU GDPR and AI Systems: What Issuers Need to Know | LinkedIn n.d. https://www.linkedin.com/pulse/eu-gdpr-ai-systems-what-issuers-need-know-jonas-frederiksen/ (accessed July 19, 2023).

[112] Legislation. Findata n.d. https://findata.fi/en/services-and-instructions/legislation/ (accessed July 19, 2023).

[113] Radley-Gardner O, Beale H, Zimmermann R, editors. Fundamental Texts On European Private Law. Hart Publishing; 2016. https://doi.org/10.5040/9781782258674.

[114] Rights (OCR) O for C. Health Information Privacy. HHSGov 2021. https://www.hhs.gov/hipaa/index.html (accessed July 19, 2023).

[115] Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022;28:31–8. https://doi.org/10.1038/s41591-021-01614-0.

[116] What's Next For AI In Healthcare In 2023. Med Futur 2023. https://medicalfuturist.com/whats-next-for-ai-in-healthcare-in-2023/ (accessed July 26, 2023).