FEDERATED LARGE LANGUAGE MODELS IN HEALTHCARE: A SYSTEMATIC REVIEW, OPPORTUNITIES AND CHALLENGES

Leon de Franca Nascimento
Institute of Computer Science
University of Tartu
Tartu, Estonia
leon.nascimento@ut.ee

Feras M. Awaysheh
Department of Computer Science
Umea University
Umea, Sweden
feras.awaysheh@umu.se

Sadi Alawadi
Department of Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
sadi.alawadi@bth.se

Abbas Cheddad
Institute of Computer Science
University of Tartu
Tartu, Estonia
abbas.cheddad@ut.ee

Albert Y. Zomaya
School of Computer Science
University of Sydney
Sydney, Australia
albert.zomaya@sydney.edu.au

Mohsen Guizani
Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI)
Abu Dhabi, UAE
mohsen.guizani@mbzuai.ac.ae

October 13, 2025

ABSTRACT

The convergence of Federated Learning (FL) and Large Language Models (LLMs) represents a transformative opportunity in healthcare. FL allows decentralized model training across multiple institutions without sharing sensitive data, which is crucial in the privacy-sensitive domain of healthcare. Meanwhile, with their exceptional natural language processing (NLP) capabilities, Large Language Models (LLMs) have demonstrated outstanding potential in healthcare applications such as clinical documentation, decision support, and patient record analysis. Despite growing interest in FL and LLM within the healthcare sector, there remains a notable gap in the literature regarding a holistic examination of these technologies opportunities, challenges, and practical applications in the healthcare context. This systematic review synthesizes cutting-edge research and identifies gaps in recent advances in combining FL and LLMs within healthcare, outlining key opportunities and challenges. This review serves as both a synthesis of current knowledge and a roadmap for future research to enable secure, collaborative, and equitable AI-driven healthcare.

***K*eywords** Federated Learning, Large Language Models, e-Health, Systematic Review

1 Introduction

Big Data has transformed the ability of healthcare institutions to manage, organize, access, and use medical information to enhance patient outcomes. It also helps reduce healthcare costs and shorten response times in addressing challenges such as medical emergencies and global pandemics [1, 2]. Large and diverse volumes of data are generated throughout healthcare services, reflecting the inherent complexity of medical practice. These data can take the form of medical images (e.g., X-rays), time-series records (e.g., heart rate monitoring), structured data (e.g., prescriptions), or unstructured text (e.g., clinical notes), which may be analyzed independently or integrated with other data types.

Nevertheless, information derived from healthcare interactions is highly sensitive. The risk of misuse poses potential threats to patient privacy and forces hospitals to invest heavily in preventing or mitigating security breaches [3, 4]. Despite these concerns, health data remain a valuable asset for developing data-driven applications that employ machine learning to reduce costs and improve patient outcomes. A wide range of applications leverage modern machine learning techniques [5, 6, 7] across various data modalities in healthcare, though most current initiatives focus on imaging and structured data [8]. Consequently, unstructured text-based data are still underutilized compared to structured data in healthcare [9].

The emergence of Large Language Models (LLMs), such as the Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT), has spurred research into their application in healthcare. These models provide new opportunities to extract and represent clinical knowledge, often outperforming previous state-of-the-art (SOTA) approaches in tasks involving unstructured clinical text [10]. A notable example is the MedPaLM model [11], which demonstrated that language models can achieve near-human performance in controlled clinical tasks, highlighting their potential for real-world medical use.

These promising results have motivated researchers across multiple medical specialties to investigate how foundational models can enhance clinical practice, provide accurate and understandable patient guidance, and support educational materials for medical trainees [12, 13, 14]. As a relatively new research field, most recent studies have focused on evaluating the performance of pre-trained LLMs in limited contexts, such as answering standardized board examination questions [15, 16] or responding to simple health-related queries [17, 18].

The prevalence of pre-trained models in these studies largely stems from the difficulty of developing LLMs from scratch, given the high financial and computational costs, as well as the limited access to large, high-quality healthcare datasets [19]. Fries et al.[20] describe these constraints as "Dataset Debt," referring to the limitations created by restricted access to specialized data for training LLMs. In healthcare, this debt arises from issues related to medical confidentiality [21], potential data breaches [22], patient consent requirements [23], and privacy-preserving regulations such as the General Data Protection Regulation (GDPR) [24]. Together, these factors make the acquisition and use of large-scale medical datasets particularly challenging. As a result, healthcare data are often limited both in availability (few datasets exist) and accessibility (external researchers rarely gain access) [25].

Federated Learning (FL) is a potential solution to mitigate the challenges posed by Dataset Debt. FL is a machine learning paradigm designed to overcome the barriers of data centralization and sharing by enabling collaborative model training across multiple institutions without transferring sensitive data [26, 27].

Within the Federated Learning (FL) paradigm, model training occurs locally on edge devices, allowing participants to train models using their private data without sharing it with third parties. Instead of transmitting raw data, participants contribute by sending model parameters, gradients, or weights to an aggregation server that coordinates the learning process. This central party manages communication among participants, aggregates local model updates, and orchestrates the overall training procedure without accessing any private data [28].

Through this approach, FL enables collaborative and privacy-preserving machine learning without exposing sensitive information. This makes it particularly well suited for healthcare applications, as it directly addresses concerns about confidentiality, privacy, and data availability [29]. Moreover, FL can be combined with additional security measures—such as blockchain, encryption methods, and multiparty computation—to enhance system resilience against vulnerabilities and cyberattacks [30].

As in other healthcare applications, most FL-based systems in the medical domain primarily rely on structured data, including electronic health records (EHRs), patient monitoring sensor data, or medical images. Consequently, unstructured text sources, such as clinical notes, remain comparatively underutilized [31].

There is therefore both a need and an opportunity to employ privacy-preserving techniques for developing large, distributed language models. This paper seeks to identify key areas of convergence, emerging opportunities, and open challenges in applying Federated Learning (FL) to Large Language Models (LLMs) for healthcare applications. By mapping current knowledge gaps at the intersection of LLMs and FL in the medical domain, we aim to support the development of models that reduce dataset debt and minimize the risk of data breaches in healthcare settings. Accordingly, this study addresses the following research questions:

Research Questions:

RQ1. How are FL and LLMs being leveraged to enhance NLP-based healthcare applications, and what technologies facilitate their adoption?

RQ2. What are the possible contributions of FL to healthcare applications that use LLMs?

RQ3. What are the main challenges associated with applying FL to healthcare applications powered by LLMs?

1.1 Contributions

This paper contributes to the growing body of research on Federated Learning (FL) and Large Language Models (LLMs) in healthcare by providing a comprehensive review of their current applications within the e-health domain. As LLMs become increasingly prevalent, this study documents the present technologies, methods, and practices involved in applying these models to healthcare contexts.

Our findings are presented using the grounded theory (GT) approach, a qualitative research method designed to identify recurring themes and concepts within a body of knowledge [32]. We apply GT to the existing literature on FL and LLMs to uncover key topics, including how these technologies can complement one another, the barriers to their adoption, and the open opportunities that remain at their intersection.

The paper also examines the main challenges in this interdisciplinary field, emphasizing issues such as data access, scalability, privacy, trustworthiness, and the computational feasibility of federating LLMs. We aim to provide insights into both the technical and ethical considerations necessary for effectively integrating these technologies into healthcare practice.

In addition, we discuss potential future research directions for this emerging area, highlighting the need to overcome current limitations and explore novel applications of integrated FL and LLM systems. While Chen et al. [33] recently offered a detailed review of the integration between language models and FL, to the best of our knowledge, this is the first comprehensive review that jointly examines FL, LLMs, and the e-health domain (Table 1).

The paper is organized as follows: Section 2 provides background information on e-health, FL, and language models. Section 3 outlines the review methodology employed. Section 4 presents the results, followed by Section 5, which analyzes them in relation to opportunities, challenges, and research directions at the intersection of these domains. Finally, Section 6 summarizes the authors' concluding remarks.

Table 1: Previous reviews on the broader area of the study

Paper	Topic	Key Contribution	Year	FL	Health	LLM
Li et al. [34]	Federated and distributed learning applications for EHRs and structured medical data	Examines FL applications on structured medical data, identifies contemporary limitations and discusses potential innovations	2023	✓	✓	-
Nguyen et al. [35]	FL for smart healthcare	Provides a comprehensive survey on the use of FL in smart healthcare, from motivations, requirements to FL designs and applications in a wide range of healthcare domains	2021	✓	✓	-
Xu et al. [36]	FL for healthcare informatics	Summarizes the general solutions to the statistical challenges, system challenges, and privacy issues in federated learning, pointing out implications and potentials in healthcare	2019	✓	✓	-
Neveditsin et al. [37]	Language Models in Medicine	Presents the capabilities of LLMs in the medical domain, with a primary focus on clinical applications	2024	-	✓	✓
Shi et al. [38]	Large Language Models in Critical Care Medicine	Evaluates the gaps in current research and examines whether LLMs can enhance clinical decision-making and improve patient outcomes in ICUs	2024	-	✓	✓
Li et al. [39]	Large Language Models for investigating EHRs	Provides a scoping review of studies using LLMs to process EHR data	2024	-	✓	✓
Yu et al. [40]	Large Language Models in Biomedical and Health Informatics (BHI)	Provides an overview of LLM applications in BHI, highlighting their potential and addressing ethical and practical challenges. Uses bibliometric methods to quantify and visualize trends, research hotspots, and impact	2023	-	✓	✓
Chet et al. [33]	Integration of large language models and federated learning	Provides a comprehensive review of the current state of the domain of LLMs combined with FL. Discusses integration advantages, challenges and future directions	2024	✓	-	✓
Ours	Federated Learning for LLMs in Healthcare	Provides an overview of discussion topics on federated learning and language models, their intersection opportunities, and challenges to address. Uses a grounded theory approach to identify themes in current research	2025	✓	✓	✓

2 Background

2.1 E-health

The e-health field encompasses computational technologies designed to enhance healthcare services through digital innovation, aiming to improve quality of life, diagnosis, disease prevention, and treatment [41]. While e-health holds significant potential to transform modern clinical practice, its success depends heavily on the digital and health literacy of both institutions and users [42].

Technologies falling under the umbrella of e-health include Health Data Analytics, Smart Homes, Virtual Care, Telemedicine, and Assisted Living [43]. These areas leverage advances in artificial intelligence and machine learning, utilizing the wealth of data stored in Electronic Health Records (EHRs) to enable predictive modeling, decision-support systems, and risk assessment based on patient information [44].

According to Fakhkhari, Bounabat, and Kassou [43], the evolution of health informatics can be divided into three distinct waves: the Technologies for Health era (1960–1999), the e-health age (2000–2020), and the current digital health age (2021–present). The latest wave gained significant momentum during the COVID-19 pandemic, when e-health solutions played a crucial role in maintaining access to healthcare services despite restrictions on movement and in-person care [45]. These conditions accelerated the digital transformation of healthcare, leading to widespread adoption of digital tools across numerous medical specialties and establishing e-health as a cornerstone of modern practice [46, 47].

Despite its achievements, e-health continues to face criticism related to data security and system scalability. Integrating diverse healthcare agents and institutions remains a major challenge, and evidence indicates that data breaches and leaks can severely undermine the trustworthiness and reliability of e-health applications [48, 49]. Consequently, the field would greatly benefit from methods that enhance privacy, scalability, and trust—ensuring that e-health solutions remain both secure and sustainable.

2.2 Federated Learning

Federated Learning (FL) is a decentralized and collaborative approach to training machine learning models, designed to address privacy concerns in sensitive domains. It redefines the traditional workflow of machine learning by reversing the typical flow of data. Instead of sending decentralized data from participants to a central server, FL distributes model parameters from a central entity to participants, allowing them to train locally without sharing any raw data [27]. The central entity then aggregates the local model parameters, weights, or gradients to construct the final global model. This structure is particularly advantageous in healthcare, as it ensures that patient data remain private while maintaining compliance with medical confidentiality standards and privacy regulations such as the GDPR [50].

In this setup, participants—known as "clients"—retain their private datasets. All clients agree upon a common model architecture, which is shared among them for local training. After completing training on their respective data, clients return their updated model parameters to be aggregated [28]. This architecture makes FL inherently privacy-preserving by design and compatible with additional security mechanisms, including multiparty computation, encryption, authentication, and other protection methods [51, 52, 53].

FL supports a variety of client arrangements, ranging from traditional server–client structures and hierarchical configurations with intermediary aggregators to fully decentralized peer-to-peer frameworks [54]. It can also be implemented synchronously—where clients contribute updates in coordinated rounds—or asynchronously, where updates are incorporated as they become available [55].

This flexibility makes FL an ideal strategy for complex healthcare scenarios involving collaborations among hospitals, networks of decentralized patient-monitoring devices, or hybrid configurations combining both. FL has been successfully applied to medical imaging, patient monitoring, and risk evaluation tasks—domains where direct data sharing would otherwise be difficult or impossible.

2.3 Large Language Models

Large Language Models (LLMs) represent a relatively recent advancement in the field of Natural Language Processing (NLP). These models are typically sequence-to-sequence neural networks built upon the transformer architecture, which leverages the concept of attention to generate outputs that account for the contextual relationships between tokens within a text [56, 57]. Originally developed for machine translation, transformer-based models soon proved highly

effective for generating coherent text from prompts, sparking widespread interest and innovation within the research community [58].

Compared to earlier NLP models, LLMs are orders of magnitude larger. While the first transformer-based models contained hundreds of millions of parameters, modern LLMs have reached sizes in the trillions [58]. This massive scale makes training them computationally expensive and data-intensive, often exceeding the technical and financial capacities of healthcare institutions [20].

The introduction of pre-trained transformers mitigated some of these challenges by enabling developers to fine-tune existing models on domain-specific data [59, 60, 61]. This approach allows LLMs to be adapted to specialized fields with relatively modest computational resources and smaller datasets. Such applications have already emerged in diverse domains, including education, finance, and risk assessment [62, 63, 64, 65].

In healthcare, clinical notes are one of the primary sources of patient information [66]. They facilitate communication among healthcare professionals, maintain continuity of care, and support patient engagement in treatment [67]. Accordingly, LLMs have shown considerable promise in processing and understanding these unstructured data sources. Early LLMs adapted for medical contexts include BioBERT [68], ClinicalBERT [69], and BlueBERT [70]. Subsequent work, such as that by Gu et al. [61], demonstrated the advantages of fine-tuning pre-trained models for domain-specific biomedical applications, achieving notable improvements in tasks like Named Entity Recognition (NER), Relation Extraction, and Question Answering.

Despite these advancements, several ethical and practical concerns remain. The use of LLMs in healthcare raises important debates about fairness, bias, non-maleficence, transparency, and privacy [71]. Additionally, because LLMs are stochastic in nature, they are prone to "hallucinations"—the generation of inaccurate or misleading information in response to user prompts [72]. This phenomenon poses a significant barrier to the safe and reliable integration of LLMs into clinical practice [73, 74].

3 Research Methodology

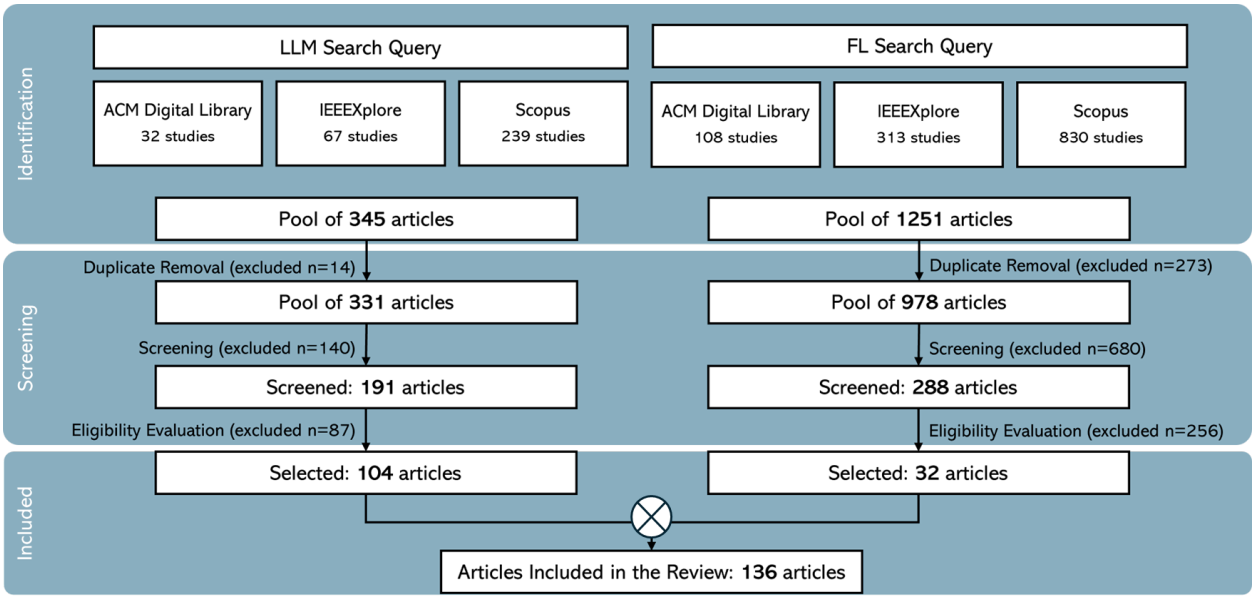

Figure 1: Systematic selection process for studies on LLMs and FL in e-health. The diagram outlines the identification, screening, and inclusion phases, detailing the sources consulted, duplicate removal, and eligibility evaluation. By applying the filtering criteria, the selected final papers were more skewed towards LLM papers.

3.1 Study Protocol

We employed the online tool *Parsifal* (https://parsif.al/) to define and implement our review protocol. The protocol consists of planning steps in which query parameters, the data extraction form, and quality standards are established. Details on the data extraction form are provided in Section 3.4.

The protocol followed the guidelines outlined by Carrera-Riveira et al. [75], utilizing the PICOC framework (Population, Intervention, Comparison, Outcome, and Context). In this study, the Population was defined as "Electronic Health Records," the Intervention as "usage of FL or LLMs," the Comparison as "Centralized Machine Learning," and the Outcomes as "efficiency and privacy-preservation aspects." The Context was "e-health applications." Research questions were derived directly from this PICOC structure, as described in Section 1.

Regarding quality standards, the following criteria were applied: the article must be available in English, and it must be peer-reviewed and published in high-quality outlets. Journals were selected based on Clarivate's Journal Citation Reports (JCR) index, requiring a quartile of Q2 or higher, while conference papers were included if they appeared in conferences rated B or higher according to the GII-GRIN-SCIE evaluation.

3.2 Information sources and search strategy

We selected papers from the IEEE Electronic Library (via IEEEXplore), the Association for Computing Machinery Digital Library, and the Scopus database, focusing on publications from the last five years. To ensure high-quality research, only studies from journals ranked JCR Q2 or higher were included. Preprints and non-peer-reviewed sources, such as arXiv, were excluded to maintain consistency and reliability, emphasizing research that had undergone formal peer review.

Two search queries were constructed to retrieve relevant publications. For the LLM portion of the review, the query was `(("Document Title": "Large Language Model" OR LLM OR GPT) AND ("Abstract": Health OR disease OR diagnosis OR hospital OR patient OR clinical))`. For the FL portion, the query was `(("Document Title": "FL") AND ("Abstract": Health OR disease OR diagnosis OR hospital OR patient OR clinical))`.

3.3 Selection Criteria

The study selection process for this review involved three steps: 1) Identification, applying the two search queries for each domain; 2) Venue quality screening; and 3) Selection criteria screening.

For the LLM domain, studies were included if they utilized an LLM and were related to healthcare applications. Studies were excluded if they were experimental or published in journals ranked below JCR Q2 or conferences below GGS B in quality assessment.

For the FL domain, inclusion criteria required that studies utilized FL, were related to healthcare applications, and worked with text- or language-based data (written or spoken). Studies were excluded if they were experimental or published in journals below JCR Q2 or conferences below GGS B.

3.4 Data Collection and Extraction

Database searches were conducted in August 2025. The authors screened titles and abstracts of 10% of all database hits (160 records) to test and refine inclusion and exclusion criteria. The remaining 90% were screened by the first author after a joint discussion of preliminary results. Data was extracted using Obsidian using a predefined extraction form (Table 2). Obsidian (https://obsidian.md/) is a note-taking software that generates links between notes. These links were used to generate tags, which aided the visualization of shared themes and topics, as well as allowing for improved retrieval of text in the synthesis phase of the study.

3.5 Data Synthesis

The final synthesis utilizes a GT approach presented in Bowers and Creamer [76]. We utilized the extraction fields defined in the prior section and, based on their content, developed preliminary coding encompassing the overall themes discussed for each category (open coding). The preliminary codes are then grouped into categories (axial coding), forming theoretical codes (selective coding), which will be the base of our discussion in the paper.

4 Results

The overall search strategy initially identified 1,596 articles, which, after screening and applying the selection criteria, were narrowed down to 136 articles (8.52%). Most exclusions occurred during the venue quality screening. For FL studies, a larger proportion of papers were excluded during selection, often because they did not involve unstructured or language-based data. Figure 1 provides a visual overview of the selection process.

Table 2: Extraction form questions used in the study to systematically collect relevant information from reviewed papers, mapping key aspects such as problem scope, proposed solutions, application domains, data usage, findings, limitations, and potential opportunities for federated learning in eHealth.

Question	Intended Information	RQ associated
What does the paper address?	Seeks to obtain information on the literature or domain problem addressed by the paper	RQ1
What is the solution proposed?	Seeks to obtain information on the overall structure of the proposed solution by the authors	RQ1
In which domain was the solution applied?	Seeks to obtain information about the medical field, task, and overall category of the solution	RQ1
What was the solution architecture?	Seeks to describe the solution proposed by the authors according to its structure and parts	RQ1
What kind of data has been used?	Seeks to describe the data types, datasets, and data selection for the studies	RQ1
What are the main findings of the study?	Seeks to highlight the main takeaways of each study regarding efficiency, quality, and application	RQ1
What are the shortcomings of the study?	Seeks to map the restrictions and limitations associated with the studies	RQ3
What are the opportunities for Federated Learning?	Seeks to identify the missed opportunities of joining FL and LLMs together for eHealth	RQ2

Analysis of publication trends showed a growing interest in these topics over time. Of the final selection, 104 articles focused on applying LLMs to e-health, while 32 examined the use of FL in healthcare applications. A complete list of selected articles is provided in the appendix, along with supplementary materials detailing each study.

Examining research contributions by country, authors affiliated with institutions in the United States, China, and the United Kingdom represented the majority (Figure 2). For visualization purposes, European Union countries were grouped together, collectively accounting for 84.16% of author affiliations. A second group contributed with the remaining 15.87% of publications with fewer selected papers composed of Canada, South Korea, Singapore, India, Israel, Australia, Japan, Switzerland, Brazil, Jordan, Palestine, Sri Lanka, UAE, Taiwan, and Norway.

Movva et al. [77] analyzed trends in LLM publications through 2023 and identified the United States and China as the most prolific contributors. They noted that academic institutions have increased publication rates, while Big Tech companies, predominantly in the United States, publish less frequently but with high impact. Industry-academia collaborations are common, but international contributions remain limited. Interestingly, our results suggest that in the subfield of LLMs applied to healthcare, Chinese institutions contribute less prominently compared to broader LLM research trends observed by Movva et al.

4.1 Grounded Theory Analysis

In the summarized notes, 350 codes were obtained from the FL subset and 1,224 codes from the LLM subset, totaling 1,574 codes during the open coding step of GT. These were grouped in the axial coding step, resulting in 11 categories for FL and 13 for LLM notes, covering challenges, opportunities, and the current application of these technologies in healthcare. For details, see the appendix, and a simplified view is shown in Figure 3.

In the selective coding step, categories were further aggregated into 8 theoretical codes (Table 3), highlighting key aspects of the systematic review to address the study's research questions. Three pillars emerged at the intersection of LLMs and FL in e-health:

1. **Data Axis** (codes 1 and 2): focuses on data accessibility and security for enabling FedLLMs.
2. **Technology Axis** (codes 3–6): addresses adoption, integration, and trust in federated LLMs for healthcare.
3. **Interaction Axis** (codes 7 and 8): examines how LLMs and FL are applied and evaluated in healthcare settings.

The interplay between these pillars highlights their interdependence. The Data Axis constrains the Technology Axis, as limited data sharing drives the need for federated approaches, influencing model performance, scalability, and trust. Integrating LLMs with FL thus requires coordination of data governance, privacy-preserving mechanisms, and distributed training protocols. Improvements in one axis, such as secure aggregation, can enable advances in another, allowing more effective fine-tuning of LLMs on distributed clinical datasets.

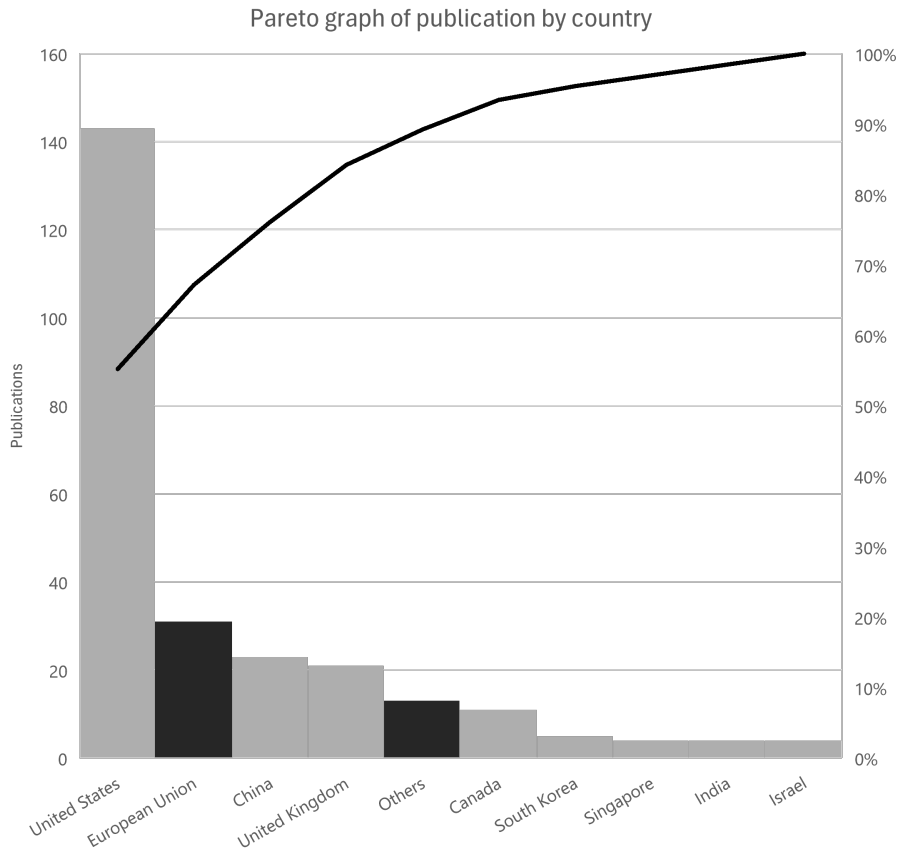

Figure 2: Pareto analysis of research publications by country, illustrating the distribution of contributions in the field. The bar chart highlights the dominant role of a few key regions, with the United States and the European Union leading in publication volume. The cumulative percentage curve underscores the concentration of research output, where a small number of countries account for the majority of contributions, reflecting global disparities in research efforts and resources.

The Interaction Axis further mediates this relationship by translating technological and data frameworks into practical use cases, including predictive diagnostics, clinical note summarization, and patient risk stratification. Insights from these applications inform refinements in model architectures, federated protocols, and evaluation metrics. This reciprocal dynamic underscores that FL and LLMs are interdependent in this scenario, and their joint optimization is crucial for effect

4.2 Current Applications

This section summarizes the main findings related to the studies encoded in the Interaction Axis, addressing the first research question of this study. We analyze common study designs observed in the studies, use cases of FL and LLMs for e-health, and the patient benefits obtained by introducing these techniques to enable healthcare applications. A brief summary can be explored in figure 4.

4.2.1 Applications of LLMs

With the adoption of generative pre-trained LLMs, healthcare professionals have increasingly explored their potential and limitations in clinical practice. A common approach in these studies is to evaluate LLMs against human standards, often using standardized board questions or benchmarking exercises. While these studies do not introduce novel computer science methods, they provide insights into the clinical capabilities of LLMs in scenarios resembling real-world medical decision-making. For example, Cai et al. [78] evaluated LLMs using 250 questions from the Basic Science and Clinical Science Self-Assessment Program, and Schubert et al. [79] tested LLMs with sample questions from a neurology board examination.

Table 3: Summary of theoretical codes and their descriptions related to LLMs and FL in e-Health.

Theoretical Code	Description	Categories
(1) Data Availability and Distribution	Describes subjects related to the access, availability, scale, and robustness of data related to EHRs usage for LLMs and FL	Data Distribution challenges in FL, Low usage of Text Data in FL, Data use limitations in FL, Challenges related to data availability in LLMs
(2) Security and Privacy Challenges	Describes subjects related to concerns pertaining to data privacy, risks of attacks, data leaks or other privacy/security related concerns	Privacy Concerns in FL, Challenges related to Data Privacy in LLMs, Security Challenges in FL
(3) Barriers to adoption	Describes subjects related to barriers pertaining to technical adoption of LLMs, like performance losses, quality or maturity or technologies	Performance and Communication challenges in FL, Challenges regarding the application of LLMs in e-health, Opportunities for e-health LLM
(4) LLM Training, Usage and Technical Aspects	Describes subjects related to the technical attributes of current usage of LLMs for e-Health, such as LLM models, training, prompting and fine-tuning	Prompt Strategies in LLMs, Federated Learning LLM training, Large Language Models Architectures, LLM architectures in FL
(5) LLM Limitations	Describes subjects related to current limitations of LLMs, like the consistency and reliability of answers, vague or general responses, hallucinations, and other limitations	Challenges related to the usability of models in LLM, Future Directions for e-Health LLMs
(6) FL opportunities in LLM for e-Health	describe subjects regarding integrating LLMs, FL, and e-Health. Examples include LLM benefits to FL setups as well as FL techniques that may improve LLMs' performance in the e-Health domain	Potential LLM applications for FL setups, Federated Learning Application for LLMs, Federated Learning Technologies
(7) Use cases and benefits	Describes subjects related to the application of LLMs or FL in the e-Health domain, as well as the patient benefits obtained through their use	Patient Benefits, LLM use cases for e-Health, FL use cases for e-Health
(8) Scientific Approaches	Describes subjects related to the methods of evaluation of LLMs and FL in the selected studies, like metrics, methods, and other	Study designs, Evaluation of LLM success

Beyond question-answering, researchers have investigated the language-processing capabilities of LLMs. Applications include extracting clinical entities, such as medications and diagnoses, from unstructured notes, enhancing mental health chatbots, and generating supportive messages for patients across various specialties. These interventions offer tangible benefits, including reduced waiting times for care [80], accelerated diagnosis [81], and faster treatment recommendations [82].

LLMs also support on-demand psychological assistance [83], helping to alleviate loneliness and emotional burden [84], and empowering patients in self-care management [85, 86]. Their text-generation capabilities have further enabled the development of educational materials across multiple medical domains, including ophthalmology [87], urology [88], orthopedics [89], and public health [90, 84, 91]. An additional key contribution is improving the accessibility and comprehensibility of health information in multiple languages [92, 93, 94]. These benefits arise from common study designs and form the basis of the primary use cases analyzed in this review.

In this section, we present the main e-health applications of LLMs, focusing on how these models leverage clinical data. We also briefly explore privacy-preserving solutions where FL has been applied to language models, highlighting the emerging intersection of these two technologies.

- **Clinical Decision Support**

  LLMs have demonstrated strong performance in answering medical knowledge questions [95, 96], which can be considered a form of decision support. Beyond simple question-answering, LLMs can provide insights

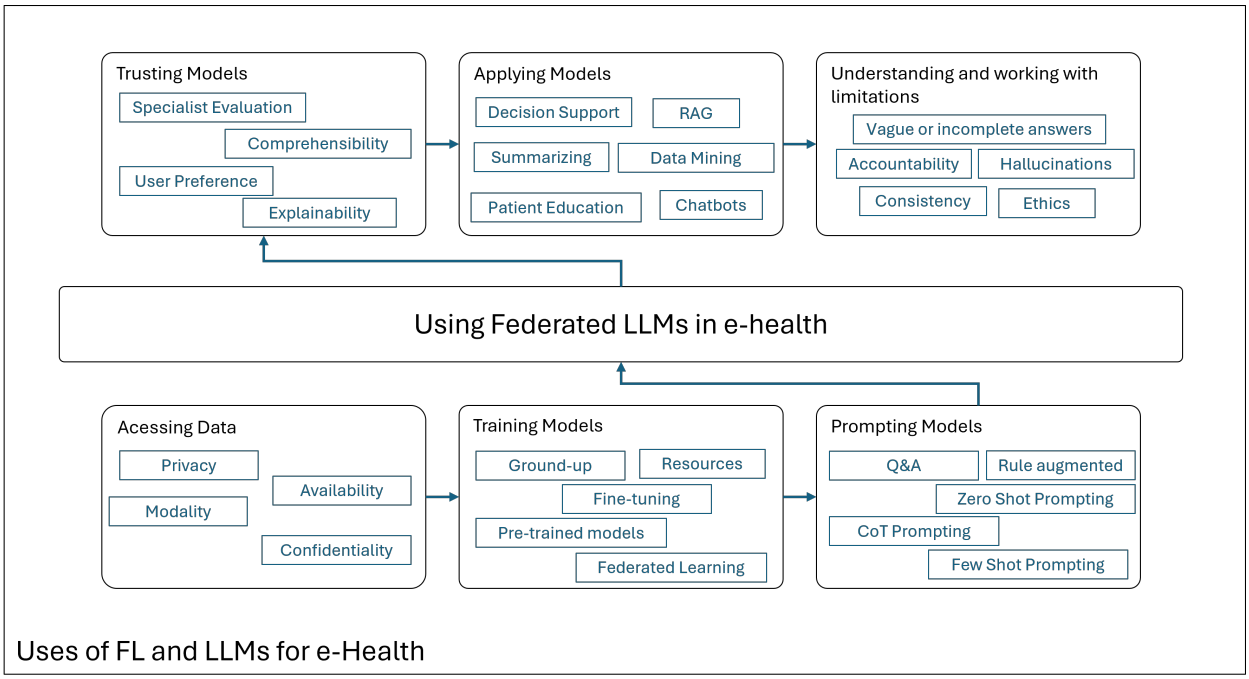

Figure 3: Overview of the applications and challenges of using Federated Large Language Models (FedLLMs) in e-health. The diagram illustrates key aspects, including data access constraints (privacy, availability, and modality), model training approaches (pre-trained models, fine-tuning, and FL), and different prompting strategies (e.g., zero-shot and chain-of-thought prompting). It also highlights trust considerations (explainability, specialist evaluation) and application areas (decision support, summarization, and patient education) while addressing limitations such as hallucinations, ethics, and accountability. This framework underscores the complexity of deploying LLMs in healthcare and the need for careful model governance.

Figure 4: Overview of LLM applications in e-Health. Key use cases of LLMs in healthcare include NER, information retrieval, data mining, text classification, chatbots, and voice assistants. Additional applications include feature extraction, decision support, data augmentation, text translation, model distillation, medical question answering (Q&A), and data summarization. These capabilities highlight the potential of LLMs in enhancing clinical decision-making, patient interaction, and biomedical data processing.

into complex clinical queries through fine-tuned models or advanced prompting strategies that enable more comprehensive evaluations.

Evidence suggests that LLM-enhanced diagnostic tools can expand the differential diagnosis options available to clinicians, reducing the risk of oversight inherent in human assessment alone [97, 98, 99]. For example, Benary et al. [100] evaluated four state-of-the-art LLMs to interpret cancer biomarker data and suggest personalized treatment strategies. Although results were modest, the study illustrates that LLMs can provide a richer set of options for physicians to consider. Similarly, Zhang et al. [101] explored LLMs for generating rehabilitation prescriptions for simulated stroke patients, leveraging patient data to propose individualized recovery plans. GPT models were able to produce clinically sound recommendations in selected scenarios, highlighting their potential to aid physicians in real-world decision-making.

Civettini et al. [102] applied LLMs to estimate transplant-related mortality and predict patient outcomes for stem cell transplantation. Their results were congruent with responses from medical residents under Cohen's Kappa measure, demonstrating that LLMs can serve as a reliable aid in clinical assessment. These approaches move beyond simple question-answering by optimizing models to generate relevant clinical content prior to deployment. For instance, Tan et al. [103] fine-tuned a GatorTron model on radiology reports and applied sentence permutation preprocessing, enabling the model to classify cancer disease response more accurately from real imaging data.

- **Risk Classification**

  While structured healthcare data has been extensively explored, free-text clinical notes often contain additional information that can be lost when data is structured. Beaulieu-Jones et al. [104] illustrate this concept by using the Clinical Longformer model to predict seizure recurrence. They found that unstructured notes encoded signals that improved classification accuracy. Similarly, Alsentzer et al. [105] employed Flan-T5 for phenotyping patients from clinical notes, identifying post-partum hemorrhage risk and detecting 47% more patients than other state-of-the-art approaches.

  Despite being primarily language models, LLMs can demonstrate strong predictive capabilities. Li et al. [106] used LLMs to identify early signs of Alzheimer's Disease from clinical notes and synthetic data, aiding early diagnosis. In more specialized areas, LLMs have been applied to Cerebral Cavernous Malformation (CCM) risk identification [107] through genetic data encoded with BERT embeddings, and to ophthalmology, supporting diabetic retinopathy screening based on features extracted from clinical notes [81].

  Emergency departments can also benefit from LLM-assisted triage. Franc et al. [108] applied ChatGPT to the Canadian Triage and Acuity Scale (CTAS) in a simulated setting, helping prioritize patients based on symptom severity. These results demonstrate that base-model performance can be significantly influenced by prompt design, enabling personalized medicine. For example, Yuan et al. [109] used carefully crafted prompts to match patients with clinical trials. The LLM parsed trial selection criteria and processed EHRs to identify eligible participants, illustrating how LLMs can bridge unstructured clinical data and patient-specific interventions.

- **Summarization**

  LLMs can indirectly support healthcare professionals by summarizing information from medical reports. Wu et al. [110] introduced a prompting method using the GLM transformer architecture to generate structured clinical findings from verbose radiology reports. Human evaluators rated the LLM-generated summaries as higher in quality and more consistent than human-generated summaries, demonstrating their potential for improving clinical decision-making.

  Another common application is simplifying medical documents, such as Randomized Controlled Trials (RCTs). Kartchner et al. [111] used a zero-shot approach with ChatGPT to extract and analyze RCT data, showing that the model could accurately retrieve information and detect missing content. Tang et al. [112] evaluated GPT models for summarizing technical reports across six medical domains, finding the outputs sufficiently comprehensive, though occasionally omitting critical details. These approaches help synthesize medical evidence for easier clinical interpretation.

  Automated summarization also enhances medical documentation. Giorgi et al. [113] developed a tool to generate clinical notes from physician-patient dialogues, reducing administrative burden and freeing clinicians to focus on patient care. Chuang et al. [114] demonstrated that engineered prompts can condense complex patient histories into manageable formats, aiding decision-making. Multi-stage summarization techniques [115] further improve information condensation across multiple documents and data sources.

  Beyond summarization, LLMs support information extraction through Named Entity Recognition (NER). Cenijk et al. [116] applied LLMs for NER in clinical nutrition research, mapping cause-effect relationships by

linking key medical concepts. Such applications enhance information retrieval in fields like clinical genetics and medical research, enabling more efficient processing of large volumes of clinical data [117, 118].

- **Patient Communication**

  Beyond diagnosis, LLMs have been applied to improve patient understanding and engagement. For instance, they have been used to generate informed consent forms [93], enhancing clarity for patients participating in clinical trials. GPT models have also assisted in rephrasing and augmenting clinical trial eligibility criteria [109], making complex conditions more accessible and understandable.

  LLMs are further employed in public health communication, generating health messages [85] and pro-vaccination content [90]. These applications enable the creation of scientifically grounded, easy-to-read messages that improve patient comprehension and agency, helping to counter misinformation and vaccine hesitancy.

  Improving comprehensibility also supports transparency in LLM-driven clinical decisions. Yang et al. [119] evaluated explainability in mental health tasks using prompt-engineered LLMs, finding that InstructGPT outperformed other models in providing interpretable outputs. Similarly, Mazumdar et al. [120] fine-tuned a GPT-3 backbone to generate explanations for its decisions, offering a means to reduce the "black-box" nature of LLMs and increase accountability in clinical applications.

- **Other Use Cases**

  LLMs have demonstrated considerable potential across diverse healthcare applications, benefiting patients, providers, and researchers alike. One notable use is providing health information directly to patients, which empowers individuals to make informed decisions. However, the accuracy and reliability of responses can vary. O'Hagan et al. [121] evaluated ChatGPT's ability to provide information on Alopecia Areata, finding that while responses were largely accurate, agreement with specialists was only moderate, indicating a need for further refinement to meet clinical standards.

  LLMs have also been valuable in mental health care, where conversational agents provide continuous support. Ma et al. [83] developed an LLM-based chatbot for a Reddit forum, offering 24/7 assistance and identifying critical issues such as suicidal ideation. This constant availability ensures timely interventions, contributing to improved mental health outcomes.

  In medical research and education, LLMs have been applied to generate synthetic clinical data and scenarios. Li et al. [106] showed that LLM-generated synthetic data can mimic real-world datasets, enabling diverse training scenarios while preserving patient privacy. Such approaches expand training possibilities and provide realistic clinical contexts that would otherwise be constrained by limited data availability.

4.2.2 FL Uses

From a federated learning (FL) perspective, this technology shows significant promise across healthcare domains by addressing privacy concerns while enabling collaborative data use. FL's design inherently preserves privacy, effectively mitigating many of the limitations associated with sharing sensitive healthcare data.

When evaluating FL use cases in e-health applications that leverage language-based electronic health records (EHRs), several key topics emerge. These applications demonstrate how FL can enhance privacy-preserving capabilities in language-focused healthcare solutions. Although some studies have begun integrating FL with large language models (LLMs), substantial opportunities remain to further exploit the combination of FL and LLMs, particularly to improve model performance while maintaining strict privacy standards.

- **Privacy-Preserving Medical Decision Support**

  In medical imaging and decision support, Chen et al. [122] employed a multimodal FL LLM to generate reports from X-ray images. This approach enables hospitals and research centers to share insights without exchanging sensitive images. The study also introduced a novel aggregation algorithm, `FedSW`, which weighs contributions based on performance prior to aggregation, showcasing a noteworthy application of federated LLMs.

  FL is particularly valuable in developing dynamic treatment regime (DTR) systems [64], where patient-specific treatments are adapted over time using continuous monitoring. It has also been applied in cardiac event prediction, analyzing distributed heart data to anticipate events such as arrhythmia or heart attacks. By pooling knowledge from multiple institutions while preserving privacy, FL enhances early warning systems and preventive care strategies. Similarly, EHR analysis [123] through FL has helped identify patterns in patient histories, improving understanding of treatment outcomes, disease progression, and potential complications.

In specialized applications, FL has been used for violence risk assessment [124], analyzing EHR notes securely from multiple sources to predict patient risk and support safety management. FL also improves adverse drug reaction (ADR) prediction [125], enabling collaborative drug safety analysis while maintaining privacy.

Neurodegenerative disease detection has benefited from FL, with applications in Parkinson's Disease [126] and early-stage Alzheimer's detection [127]. Models trained on distributed datasets can identify subtle behavioral or imaging patterns, improving early diagnostic accuracy without compromising patient privacy.

Finally, FL addresses challenges in EHR-based model training by avoiding data movement that could expose sensitive information. Liu et al. [64] demonstrated this by using FL to analyze historical patient data across multiple providers, extracting features from EHRs to predict tumor treatment outcomes and patient trajectories securely.

- **Privacy-Preserving Patient Support**

  Mental health is an area where FL can have a significant impact. Shin et al. [128] introduced FedTherapist, a system providing psychological support using multimodal patient data (text and speech) from smartphones within a federated framework. Their method aggregates additional sensitive information, such as geospatial data, to contextualize language-based content using a Context Aware Language Learning (CALL) approach. This enables analysis and identification of sensitive psychological profiles across multiple participants while preserving the confidentiality of individual records.

  FL has also been applied to detect mental health distress without compromising privacy. Ahmed et al. [129] developed a federated NLP approach to identify depression symptoms from text by encoding hypergraphs derived from participants' mobile phone data. Such approaches allow the creation of more robust and generalized models, offering personalized interventions while ensuring patient privacy.

4.3 Technologies

In this section, we explore the technologies and methods currently used to enable e-health applications powered by LLMs. We focus on LLM architectures, the use of prompt engineering, and the evaluation methods employed to assess LLM performance. Our review found that FL studies generally did not propose new LLM methods. Instead, they applied established techniques, such as FedAvg, or focused on deploying FL in healthcare settings rather than introducing novel methods. The topics discussed in this section primarily relate to codes 4 and 5 from the GT analysis.

4.3.1 LLMs Currently in Use for Healthcare

Analysis of the selected papers revealed a diverse range of neural network architectures, as shown in Figure 5. Among these, OpenAI's GPT models were the most commonly used in healthcare applications. Other popular families include BERT-type models, LLaMA-type models, and several less common architectures. While a variety of model architectures are employed, smaller pre-trained models are more frequently used in healthcare settings due to resource and data constraints. In this section, we examine the trends and characteristics of these architectures, their adoption patterns, and their suitability for different healthcare applications.

- **GPT-based Models**

  ChatGPT and its iterations, including ChatGPT 3.5 and ChatGPT-4, have been widely evaluated for clinical decision support, patient interaction, and data summarization [111, 130, 131, 132, 133, 117]. Their popularity is likely due to the user-friendly front-end, which allows adoption by non-technical healthcare personnel, including physicians and nurses [87, 81, 134, 79, 92, 135].

  While some studies used earlier GPT versions such as GPT-2 [136, 137], most research focused on GPT-3 and GPT-3.5 turbo, leveraging their efficiency in processing unstructured text, including dialogues, questions, and clinical notes [138, 139, 140, 141]. Most studies relied on OpenAI's pre-trained weights, which limits domain-specific performance evaluation. Microsoft's Bing Chat, also GPT-based, was similarly employed in studies with parallel objectives [142, 143, 88].

  The introduction of GPT-4 brought improved conversational capabilities and context comprehension. GPT-4 has demonstrated utility in identifying eligible patients for clinical trials [144], performing healthcare economic modeling [145], extracting patient chief complaints [146], and generating synthetic datasets for model training [106], highlighting its versatility across complex healthcare tasks.

  Domain-specific fine-tuned GPT variants have also emerged. InstructGPT has been used to generate distilled data annotations for entity extraction [147] and to improve decision explainability in mental health scenarios [119]. GPTFX [120] provides post-hoc explanations of GPT decisions, further supporting explainable AI. Other

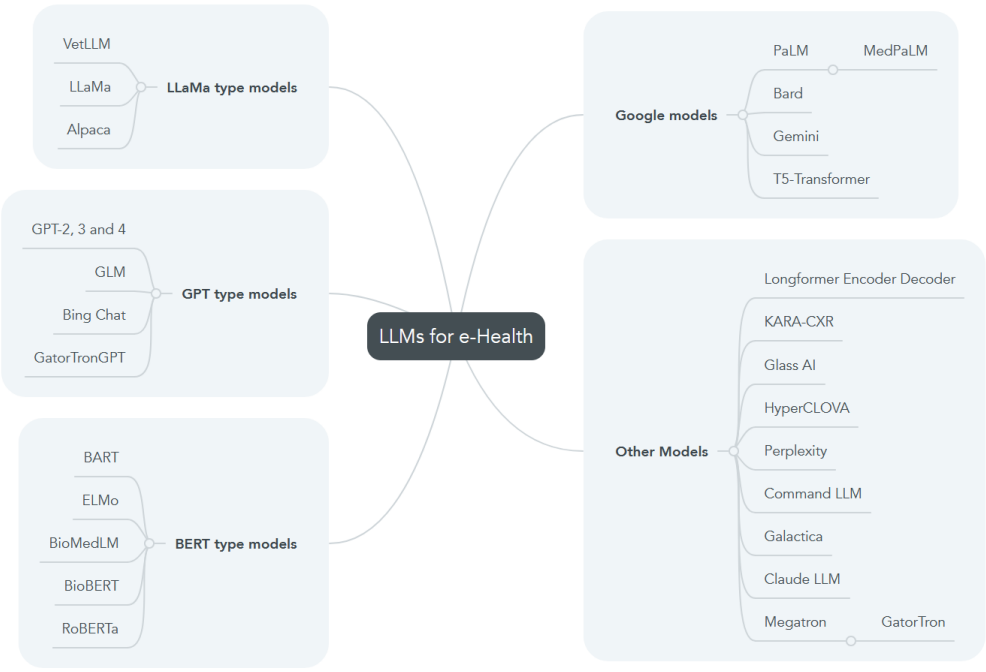

Figure 5: Categorization of LLMs used in e-Health applications. The diagram classifies LLMs into different families, including LLaMa-type models (VetLLM, LLaMa, Alpaca), GPT-type models (GPT-2, 3, 4, GLM, Bing Chat, GatorTronGPT), BERT-type models (BART, ELMo, BioMedLM, BioBERT, RoBERTa), Google models (PaLM, MedPaLM, Bard, Gemini, T5-Transformer), and other specialized models (Longformer Encoder Decoder, KARA-CXR, Glass AI, HyperCLOVA, Perplexity, Command LLM, Galactica, Claude LLM, Megatron, GatorTron). This categorization highlights the diversity of LLM architectures applied in healthcare research and practice.

domain-trained GPT models include GatorTronGPT [148] (trained on clinical notes from the University of Florida), BioGPT [146], PubmedGPT [149] (trained on PUBMED), and PhenoGPT [150] (using BioLarkGSC+ data).

- **BERT-based Models**

  The second most common LLM family in healthcare applications is based on BERT. Due to its relatively smaller size, BERT is more cost-effective for domain-specific fine-tuning compared to larger models. Variants such as base BERT, RoBERTa, and ClinicalBERT have been applied to benchmark various medical NLP tasks [149, 151]. For instance, Tan et al. [103] compared multiple transformer models, including BERT, BioBERT, BioClinicalBERT, BioMegatron, DeBERTa, GatorTron, PubMedGPT, RadBERT, RoBERTa, and XLNet, highlighting their relative performance on specialized healthcare NLP tasks.

  Fine-tuning on clinical datasets enables these models to handle domain-specific terminology effectively [141, 152, 150, 137]. Li et al. [106] further demonstrated the integration of ClinicalBERT with memory networks (Mem) to improve text retrieval and classification, enhancing the model's ability to manage long-document context, which is critical for EHR narratives.

  Domain-specific adaptations, such as the SAFFRON model [116], leveraged BERT, RoBERTa, and BioBERT to generate cause-effect graphs from clinical NLP tasks, exemplifying the flexibility of foundational models for text extraction and relation analysis. Similarly, BioBERT embeddings have been used for EHR text vectorization, enabling semantic extraction [107], privacy-preserving applications [109], mental health topic mapping [153], keyword extraction from speech [136], and patient-trial matching [109].

  BERT networks can also support other LLMs in hybrid workflows. Meoni et al. [147] used InstructGPT-3 to generate annotations for unlabeled clinical notes, which were then used to fine-tune BERT, improving classification and annotation in specialized datasets. Moreover, BERT-based models have facilitated model distillation; for example, RoBERTa was used to train smaller TinyBERT models for NER tasks [118].

- **Google Bard/Gemini, T5, and PaLM Models**

Google's family of models, including T5, LaMDA, PaLM, and Gemini, has been applied across a variety of healthcare NLP tasks. T5 and its fine-tuned variant Flan-T5 have been used for summarizing patient complaints [154], extracting features from clinical notes [155], and detecting suicide ideation from social media text [152]. Alsentzer et al. [105] demonstrated that Flan-T5 can identify Post-Partum Hemorrhage cases from clinical notes and classify subtypes with minimal intervention, highlighting its potential for domain-specific concept extraction.

Bard (LaMDA) and Gemini have primarily been explored in QA applications. Examples include ophthalmology question-answering [87, 156] and assessments of race-based biases in medical calculations [157]. Tsoutsanis et al. [95] compared Bard with ChatGPT, LLaMA, and BingChat in the Multi-Speciality Recruitment Assessment test, illustrating the utility of chat-based interfaces in clinical problem-solving.

PaLM has demonstrated scalability across large medical datasets. Singhal et al. [96] evaluated the base 540B-parameter model on medical tasks and fine-tuned it on Medical QA datasets (Flan-PaLM), achieving SOTA performance, though reasoning flaws remained. This motivated the development of MedPaLM, an instruction fine-tuned version that improved clinical adaptability and reduced the potential for harmful outputs [158]. Civettini et al. [102] evaluated MedPaLM on hematopoietic stem cell transplantation eligibility, finding strong concordance with medical residents' decisions, confirming its practical potential in healthcare.

- **Other Architectures**

  Several additional LLM architectures were identified in the selected studies, though less frequently than GPT, BERT, or Google models. Meta's LLaMA family, including Vicuna and Alpaca, has been applied in mental health, veterinary medicine, and report summarization tasks [159, 160, 161]. The Longformer Encoder-Decoder (LED), optimized for long-range dependencies, proved suitable for extensive clinical texts. Giorgi et al. [113] used LED to summarize patient-doctor conversations with an in-context learning approach, producing summaries comparable to those of physicians. Beaulieu-Jones et al. [104] applied LED to predict seizure occurrences from clinical notes, outperforming alternative text processing methods.

  GatorTron-based architectures excel in large-scale medical NLP tasks, particularly NER and classification, leveraging extensive datasets from the University of Florida [162]. Other architectures from companies and research institutions, such as ChatGLM, HyperCLOVA, and Glass AI, have been applied to language generation, information retrieval, multilingual mental health applications, and neuroimaging [110, 83, 163]. Glass AI produces consistent reports in clinical settings, though not surpassing physician accuracy, demonstrating utility in decision support.

  Specialized models, including BioMedLM, Perplexity, and Galactica, have shown success in clinical data processing and augmentation, improving downstream model performance [164, 100, 165]. BART has been explored for text generation and clinical report summarization [166]. Additional architectures, such as Claude-instant-v1.0, Command-xlarge-nightly, and Bloom, demonstrate varying strengths in speed, accuracy, and multilingual capabilities, with Bloom noted for handling diverse patient populations [88, 167, 85].

4.3.2 Privacy-Preserving LLM and related architectures under FL

For the use of transformer- and LLM-based architectures on EHRs under FL, a smaller set of models and methods was observed compared to purely LLM-focused studies. These ranged from traditional sequence models like LSTMs to larger transformer-based models, including LLaMA 7B, highlighting challenges such as client resource heterogeneity, configuration, and the relatively recent adoption of Federated LLMs.

Simpler architectures remain relevant in specific domains. For instance, Ahmed et al. [129] explored mental health states from text-based EHR data using uni- and bidirectional LSTMs. Inputs were vectorized via WordNet graph embeddings and processed through an attention mechanism—an approach inspired by Transformers—achieving an AUC of 0.86 for PHQ-9 depression classification. Similarly, Sarlas et al. [126] applied an adversarial auto-encoder to audio data for Parkinson's disease, preserving patient privacy while capturing complex data profiles. These studies demonstrate opportunities to integrate traditional architectures with privacy-preserving frameworks.

FL studies have increasingly leveraged transformers like BERT. Shin et al. [128] explored three approaches: (a) Fixed-BERT + MLP, where pre-trained embeddings were fine-tuned for downstream tasks; (b) End-to-End BERT + MLP, allowing backpropagation across the entire model for improved EHR adaptation; and (c) a LLaMA-7B setup, though deployment to smartphones was constrained by model requirements. Lightweight variants such as DistilBERT and Mobile-BERT also showed promise for mobile health applications due to their reduced resource footprint.

Other studies illustrate multimodal applications. Lu et al. [168] trained a BERT-BASE backbone for the FedMedVLP model, integrating image-to-text and text-to-image retrieval with medical Q&A tasks, outperforming other SOTA

Figure 6: Prompt engineering methods used in the studies. This schematic categorizes key techniques, from in-context learning approaches like zero-shot and few-shot prompting to advanced reasoning methods such as Chain-of-Thought (CoT) and diagnostic prompting.

models. Chen et al. [122] trained a transformer-based report generator from X-ray image features in a federated setting, achieving higher BLEU-1 scores than individual local models. These approaches highlight the potential of FL to enable collaborative training of LLMs in healthcare while preserving data privacy and improving model performance.

4.3.3 Prompt Engineering

Prompt engineering is a fundamental component of LLM use in healthcare, as the model's output is highly dependent on how tasks are framed and instructed. The choice of prompting strategy directly impacts the relevance, accuracy, and usability of the results in clinical and research settings. Broadly, prompting techniques can be categorized into three types:

Zero-shot prompts, where the LLM is expected to perform a task using only the instructions provided, without additional examples; **Few-shot prompts**, in which the model receives instructions along with a small set of example input/output pairs [169]; and **Chain-of-Thought (CoT) prompts**, where the LLM is instructed to reason through a problem step by step, producing a structured explanation or reasoning process to reach the answer [170].

- **Zero and Few-Shot Prompts**

  Zero-shot prompting is widely used in medical Q&A studies, where LLMs are tasked with answering questions or performing direct tasks such as information extraction, summarization, or classification without prior examples. Kartchner et al.[111] evaluated ChatGPT and a fine-tuned GPT model (GPT-JT) to extract key data from Randomized Clinical Trials (RCTs) in a zero-shot setting. Similarly, Bhate et al.[139] used zero-shot prompting to extract social determinants from clinical notes, achieving promising results. For summarization, Tang et al.[112] used ChatGPT to generate abstracts of systematic reviews, finding high extractiveness but low BLEU scores compared to human summaries.

  Zero-shot methods were also applied to multimodal and veterinary tasks. Lee et al.[130] prompted ChatGPT and the domain-specific KARA-CXR model to describe findings from DICOM X-Ray data, with KARA-CXR outperforming ChatGPT-4. Jiang et al.[159] applied zero-shot and fine-tuned Alpaca models to veterinary notes, finding that fine-tuning improved diagnosis rates, though zero-shot outputs were reasonably accurate. On content generation, Kianian et al.[87] and Karinshak et al.[90] demonstrated that simple zero-shot prompts could generate understandable health messages for the public.

  Few-shot prompting enhances performance by providing sample input-output pairs. Agrawal et al.[171] found that adding one example improved GPT-3's accuracy on tasks like co-reference resolution, medication

extraction, and status classification. Shyr et al.[151] showed that few-shot ChatGPT outperformed zero-shot for identifying rare diseases, despite BioClinicalBERT performing better overall. Jo et al.[84] used few-shot prompting with HyperCLOVA on dialog-based data to provide mental health support, reducing professional burden but producing open-ended outputs. Giorgi et al.[113] demonstrated that in-context few-shot examples improved GPT-4 and LED model summaries of patient-provider conversations. Nair et al.[115] applied a multi-stage few-shot approach with GPT-3 for dialogue summarization, finding optimal results at three-shot prompting, though additional contextual examples did not improve performance.

Overall, zero-shot prompting enables emergent behavior, while few-shot and in-context learning enhance output accuracy and relevance in complex clinical tasks.

- **Structured Prompts**

  Structuring prompts—both instructions and expected outputs—can improve the quality and relevance of LLM outputs [172]. Lim et al.[85] used structured prompts with the Bloom model to generate health messages based on frequently tweeted content. The structured prompts produced outputs perceived as clearer and higher quality than typical human tweets. Similarly, Benary et al.[100] explored structured prompts with BiomedLM, Perplexity, ChatGPT, and Galactica to provide targeted cancer therapy suggestions for fictional patients, finding that prompt structure enabled a wider variety of responses than human suggestions, acting as a potential support tool for oncologists.

  Bernstein et al.[173] applied structured prompts adhering to the P.E.A.R.L.S guidelines to make responses more empathetic, and independent evaluators could distinguish LLM-generated responses from human responses only 61% of the time. Shyr et al.[151] compared structured versus unstructured prompts for ChatGPT, finding that simpler prompts sometimes performed better, likely due to ChatGPT's conversational optimization.

  Structured prompts are also used in sequential and reverse-prompting strategies. Panagoulias et al.[82] combined three sequential prompts with rule-based reasoning to mimic physician diagnostic processes. Wu et al.[110] used reverse prompting with fine-tuned models to generate richer clinical findings from summarized radiology notes, supporting professional training in lesion identification.

  Finally, prompting for uncertainty is critical in sensitive healthcare contexts. Civettini et al.[102] uniquely allowed models to respond with "I don't know," reducing the risk of hallucinations and improving reliability in clinical scenarios.

- **Chain-of-Thought Prompting and Specialized Derivatives**

  Chain-of-Thought (CoT) prompting improves the interpretability and trustworthiness of LLM outputs in healthcare by explicitly showing intermediate reasoning steps [174]. Chen et al. [160] applied CoT prompting to identify cognitive distortions in therapist-annotated speech (Therapist QA Dataset). While CoT improved ChatGPT-3.5 performance on zero-shot tasks, GPT-4 showed no improvement. They further developed a specialized method, Diagnosis-of-Thought (DoT), which combined subjectivity assessment, contrastive reasoning, and an integrative evaluation. DoT outperformed standard CoT, enhancing classification for both GPT-3.5 and GPT-4.

  Savage et al. [175] introduced "diagnostic reasoning" CoT prompts, designed to mimic physicians' thought processes, including differential diagnosis, Bayesian inference, and analytical reasoning. Using one-shot examples, this approach modestly improved GPT-3.5 and GPT-4 diagnostic performance while making outputs more interpretable for clinical use.

  Yang et al. [119] proposed "Emotion-Enhanced Chain-of-Thought," augmenting CoT with few-shot examples of emotion classification for mental health analysis. Although BERT-based fine-tuned models still outperformed it, this method achieved the best performance among autoregressive models. The study noted that outputs were highly sensitive to prompt phrasing, highlighting a limitation in stability for such reasoning-based approaches.

4.3.4 Evaluation of LLMs

When applied to the healthcare domain, there is a need for the definition of metrics to evaluate the success of a model in its purpose task. As language models, we've discussed that the use cases of LLMs span diverse domains, from classification, extraction, and health message generation. As such, the metrics used for evaluating the performance of LLMS reflect the diversity of tasks in which these models are used. A brief summary can be seen in figure 7

- **Classification Metrics**

  Since some of the tasks in the studies related to identifying relevant word classes, extracting medical information from raw text, and, naturally, classifying patients into groups, traditional classification metrics,

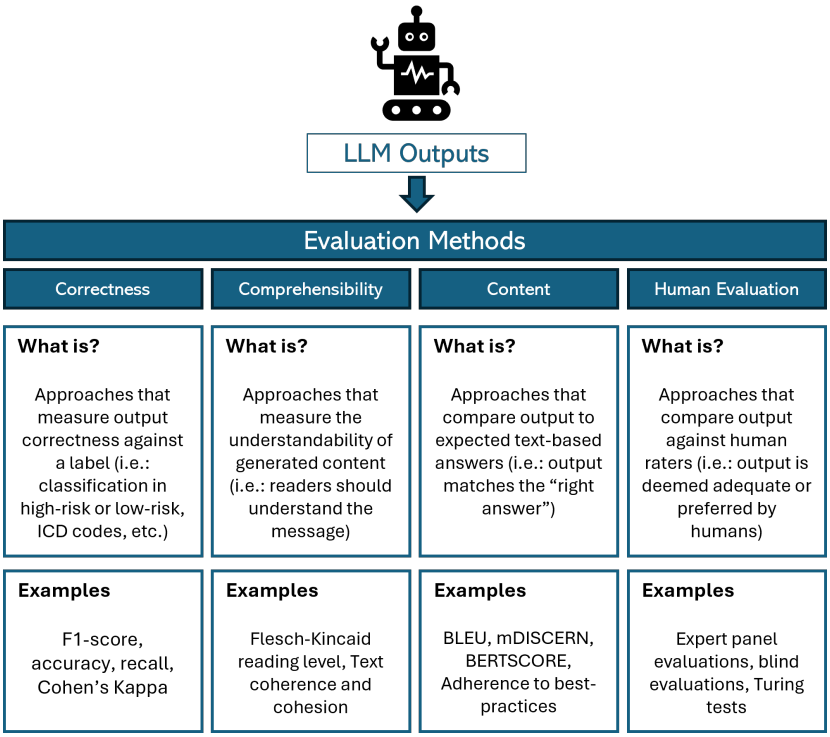

Figure 7: Common evaluation methods for assessing the outputs of LLMs. The evaluation approaches are divided into four key dimensions: Correctness, Comprehensibility, Content, and Human Evaluation. Each category defines its purpose and provides examples of commonly used metrics.

like accuracy and precision, were often present in the papers. Some examples include correct diagnosis classification with F1 score [159, 99], extracted terms accuracy in NER [151], question answer accuracy [132], triaged patients correctness [108], risk prediction using the ROC curve and AUC [104, 176], and precision and recall of recommendations [100].

Another common evaluation method was identifying the agreement of models with expected answers, be it from medical consensus, standardized results, or professional opinion. For this purpose, Cohen's Kappa was often used in papers. Usage to measure the distilled models' agreement with the teacher models [118] and agreement among consensus and physicians [81] illustrate the use of this metric in LLM evaluation.

- **Comprehensibility**

  Besides the correctness of outputs generated by the models, authors also often evaluated the understandability and ease of reading LLM-generated responses. As language models, the community is particularly interested in answers provided by these models that are easy to understand by patients, who may engage with these tools through chatbot interfaces like ChatGPT. Another motivation is constructing LLM-based tools that enable high-quality patient interaction with minimal provider intervention in digital platforms, social media, or medical assistants.

  Among the main methods of evaluating comprehensibility is the Flesch-Kincaid ease of reading and reading grade formulas [177]. It has been used to measure the readability of surgical consent forms rewritten by LLMs [93], to evaluate radiologist and AI summaries on x-ray reports [92], with LLM provided info on shoulder surgery [89], rheumatology questions [143], uveitis questions [87], and other applications. Beyond Flesch-Kincaid, authors were also concerned about the coherence and comprehensiveness of LLM-generated summaries [112] and the quality and clarity of AI-generated messages [134], often employing a panel of raters to rate the quality of the texts generated.

- **Content**

  LLM outputs in healthcare are typically evaluated based on how closely they align with expected or clinically acceptable text, using both automated and human-centered approaches.

Automated Text-Based Metrics. BLEU [178] and ROUGE [179] measure n-gram overlap between model outputs and reference text. For example, Wu et al. [110] used BLEU to evaluate LLM-generated medical summaries, while Mazumdar et al. [120] applied ROUGE to assess post-hoc explanations from GPTFX. BERTScore [180] improves on these by using contextual embeddings to measure token similarity; Islam et al. [146] applied it to evaluate autocompleted patient chief complaints. METEOR [181] uses an F1-based approach for n-grams and was employed by Tang et al. [112] to evaluate medical summaries.

Health Information Quality Metrics. DISCERN [182, 183] and its shortened form mDISCERN [167] assess the quality and trustworthiness of health content. Hurley et al. [89] used DISCERN and JAMA Benchmark Criteria to evaluate ChatGPT outputs in orthopedic surgery, finding generally low scores, indicating limitations in LLM reliability. These metrics help ensure LLM outputs provide accurate, well-referenced information.

Expert Evaluation of Relevance and Appropriateness. Several studies rely on panels of medical professionals to assess content clarity, specificity, safety, and appropriateness. Abi-Rafeh et al. [97] evaluated Bard's post-operative care instructions, while Lahat et al. [135] assessed the clarity and originality of LLM-generated research questions in gastroenterology. O'Hagan et al. [121] and Dennstdt et al. [184] similarly quantified the suitability of LLM outputs for patient or professional use.

Adherence to Guidelines and Reference Documents. Some evaluations measure how well LLM outputs align with clinical guidelines or consensus documents. Fisch et al. [185] compared responses about meningitis to medical guidelines, Chowdhury et al. [186] assessed cataract-related answers, and other studies examined rehabilitation [101], dentistry [187, 86], and general medical information [87, 188]. These evaluations ensure that outputs are not only coherent but also clinically valid.

Overall, evaluation strategies range from automated n-gram and embedding-based metrics to expert panels and guideline adherence checks, providing complementary perspectives on LLM performance in healthcare.

- **Human Evaluation**

  Given the clinical focus of LLM applications, many studies evaluated model outputs against healthcare professionals' expertise, perceptions, and preferences. Duong et al. [189] assessed LLM performance on genetics-related inquiries by comparing AI outputs to human expert responses. Similarly, Tsoutsanis et al. [95] employed the MSRA exam format to benchmark LLMs against standards encountered by practicing physicians.

  Diagnostic accuracy and reliability were examined in various contexts. Gopalakrishnan et al. [81] compared LLM-generated diabetic retinopathy notes to expert raters, while O'Hagan et al. [121] solicited specialist feedback to evaluate the appropriateness of AI-provided patient advice. Yang et al. [119] focused on interpretability and acceptability in mental health, highlighting professional perspectives on clarity and utility in patient care.

  Some studies explored patient-facing or public health contexts. Jo et al. [84] used qualitative methods, including interviews and focus groups, to evaluate conversational AI in public health, and Ye et al. [143] studied patient trust and perceived authenticity of AI responses in rheumatology. Decker et al. [93] compared LLM- and surgeon-authored informed consent forms, emphasizing clarity and legal considerations.

  LLM outputs were also benchmarked against clinical decision-making. Wilhelm et al. [167] reported strong alignment between GPT-4 recommendations and physician assessments across multiple specialties. Sorin et al. [133] compared ChatGPT suggestions to tumor board decisions in breast cancer cases, while Civettini et al. [102] evaluated agreement with residents' decisions in hematopoietic transplant scenarios. Lim et al. [85] assessed health message generation against human-authored content, and Peng et al. [148] conducted a Turing-style test with Gatortron using real patient data to gauge AI indistinguishability from human responses.

  Finally, studies in specialized fields highlighted LLM potential in nuanced decision-making. Kottlors et al. [98] explored concordance of LLM-generated differential diagnoses with expert consensus, and Benary et al. [100] compared personalized oncology treatment suggestions from LLMs to expert recommendations. Collectively, these studies illustrate how human evaluation serves as a critical benchmark for assessing LLM reliability, clinical appropriateness, and integration potential in healthcare practice.

4.4 Challenges

While LLMs hold significant promise for advancing digital health technologies, several challenges must be addressed before they can be fully integrated into clinical practice. A primary concern is data accessibility, which raises privacy risks when handling, sharing, and analyzing sensitive health information. Limited access to large, high-quality datasets further constrains model training, making data a scarce and valuable resource. In this section, we examine theoretical codes along the technology axis, focusing on barriers to adoption and inherent LLM limitations.

Table 4: Key challenges in federated learning and corresponding mitigation strategies, highlighting approaches for privacy preservation, resource efficiency, data quality, reproducibility, trustworthiness, and the integration of language models in decentralized settings.

Challenge	Potential Mitigation Strategies
Prioritizing privacy and secure aggregation	Differential privacy, secure multi-party computation (SMPC), homomorphic encryption, secure aggregation protocols in FL, and private federated averaging.
Resource-efficient training in heterogeneous environments	Adaptive model pruning, federated optimization techniques (e.g., FedAvg, FedProx), client-side compression, low-bandwidth communication protocols, and model distillation.
Robustness to data imbalance and quality variations	Client selection strategies based on data quality, federated re-weighting, cross-device federated learning with bias mitigation, and federated data augmentation.
Reproducibility, variability, and correctness of outputs	Consistency protocols in model updates, federated validation strategies, reproducible training pipelines, and consensus aggregation for model updates across heterogeneous clients.
Ensuring trustworthiness and accountability	Blockchain for auditability, federated learning accountability frameworks, explainable federated models (XAI), model provenance tracking, and transparent client contribution recording.
Low use of text data in FL	Lightweight LLMs, LoRA, Prompt Engineering

LLMs also demand substantial computational resources for training and inference, a requirement that becomes even more pronounced in federated learning scenarios. Despite their rapid improvements, these models remain vulnerable to biases and errors, underscoring the need to carefully consider ethical and practical implications before using them for clinical decision support.

The literature demonstrates a strong interplay between technological choices and data-related constraints. For example, the selection of FL architectures and parameter-efficient fine-tuning strategies often directly addresses challenges in data accessibility, heterogeneity, and privacy. This highlights a synergy between technical innovation and careful data management in the deployment of LLMs for healthcare.

- **Privacy and Security Concerns**

  While pre-trained models provide efficiency, they are limited by data representativeness, creating a tension with privacy-preserving approaches like federated learning that require additional complexity but enable richer domain adaptation. Many authors understand that data privacy is an important factor when working with health data in these models [150, 190, 108], as well as the ethics of training LLMs on social media and smartphone interactions [85, 189, 128]. Even under the prism of FL, studies are still concerned with possible privacy risks related to data leaks [191], eavesdropping attacks [192], side-channel attacks [128], and malicious activities [193], meaning that the adoption of FL for this scenario is not a silver bullet on its own.

- **Performance and Communication Challenges**

  The FL paradigm introduces additional complexity to LLM training, as the process becomes decentralized. This brings challenges related to participants' resource efficiency, including sending and receiving updates, processing training computations, and running large-scale inference on edge devices. Ensuring fair and secure contributions in FL may introduce communication and performance overheads [194, 122, 195, 196], and models must remain resilient to skewed data distributions [168, 197, 198], heterogeneous data quality [64, 199], and potential data drift over extended participation cycles [200].

  Beyond these practical concerns, LLMs remain resource-intensive, posing technical challenges for scaling to edge devices for both training and inference. Many studies relied on pre-trained models, leveraging their encoded knowledge through prompting. While prompt engineering can yield reasonable results, open pre-trained models, such as OpenAI's ChatGPT and Google's Gemini, are trained on broad datasets and are not specialized for biomedical applications. This limits their applicability in healthcare and may explain some suboptimal results reported in the literature.

  Privacy risks further complicate the use of third-party models via chat interfaces or API requests, making them less suitable for e-health applications. To address this, some studies implemented fine-tuned pre-trained models locally, which allowed for personalized models but constrained the scale of models that could be trained. For context, while GPT-3 contains 175 billion parameters [169], the fine-tuned models reviewed ranged from 14.5M to 175B parameters, with the majority in the low billions (Table 5).

  The challenges of training large-scale LLMs are evident in the scarcity of studies training models from scratch. Notable examples include Gatortron [162] and GatortronGPT [148], derivatives of the Megatron transformer trained on clinical notes from the University of Florida, and Google's Med-PaLM models [96, 158], trained

on large-scale medical question datasets. Med-MLLM [94], a multimodal LLM for chest X-ray COVID classification trained on the MIMIC-CXR dataset, and Kara-CXR [130], trained on data from a Korean hospital, also fall in this category. These models, each with billions of parameters, require substantial computational resources and data, highlighting the need for optimized training and parameter-efficient fine-tuning techniques, such as Low-Rank Adapters (LoRA), transfer learning, and p-tuning, to facilitate deployment and scaling on edge devices.

- **Challenges Regarding Correctness and Trustworthiness**

  Even after overcoming barriers of data accessibility and resource efficiency, significant challenges remain regarding the correctness and trustworthiness of AI models in healthcare. First, LLMs are trained on historical data, which may become outdated by the time the models are deployed. Omiye et al. [157] demonstrated that popular LLMs, when prompted for technical information relying on patient demographics, produced outdated calculations no longer valid for race-based measures, such as renal function. This could lead practitioners or automated agents to make decisions based on incorrect data, potentially harming patients. Similar issues were reported in orthopedic management [142] and dentistry [187].

  This limitation is linked to the fact that LLMs do not make the same inferences as practitioners and cannot consult original knowledge sources. For instance, Wilhelm et al. [167] evaluated LLM-generated health messages, finding that while the content was well-written, the models failed to provide reliable references. All evaluated models scored zero in the JAMA benchmark, which assesses authorship, attribution, disclosure, and currency. Karinshak et al. [90] reported similar findings in pro-vaccination campaign messages: although human raters preferred AI-generated messages to CDC materials, they judged AI outputs as less trustworthy than messages authored by medical professionals.

  These issues are partly due to the stochastic nature of LLMs [72]. Outputs are generated based on statistical relationships between contextual embeddings of tokens in the training data. Consequently, models are non-deterministic: repeated prompts can yield different outputs. Franc et al. [108] observed that in a triage scale task with 30 replicates, repeatability (same prompt, different results) accounted for 21% of variation, while reproducibility (similar prompt, different results) contributed 4%. Variability was also reported by Sosa et al. [142], where queries sometimes failed to provide complete guidance, and by Beaulieu-Jones [190], where answers to close-ended questions varied in approximately one-third of prompts.

- **Low availability of text data in FL**

  In the introduction of this paper, we discussed the diversity of data types generated in healthcare practice. Patient-provider interaction generates a wealth of information through measurements, questionnaires, form responses, images, and clinical notes. We found a significantly large pool of relevant articles when querying the databases for studies related to language models and FL in e-health. However, especially for FL, the majority of papers (97%) are associated with data that is not text-based, which reinforces the concept that clinical notes are still an underutilized resource of information for healthcare AI models. Even for the LLM papers, text dataset availability was not based on large consolidated datasets (except for MIMIC-IV [201]), with models often being trained on local private datasets [162, 82, 104], scrapped social media data [189, 153, 202], question banks [96, 132, 79] or authored questions [121, 157], patient vignettes [108, 100], dialog datasets [115, 203, 166, 113].

  These examples corroborate the concept of dataset debt presented by Fries et al. [20], highlighting that the need to have large-scale datasets may not be addressed with publicly available data. In the papers selected for this review, authors often remarked limitations on the fragmentation of NLP datasets [171], access to relevant clinical data [151], small sample sizes [115, 131], and biases related to data distribution being limited to single institutions [130, 103]. Authors also often needed to rely on simulated conditions [97, 185, 100], which may introduce biases and not represent real scenarios faced by real physicians and other practitioners.

- **Stochasticity, Hallucinations, and Explainability Challenges**

  Another critical issue related to the stochastic nature of LLMs is the potential introduction of incorrect or incomplete information. Since end tokens are among the possible outputs, models may terminate a chain of reasoning without addressing all points necessary to satisfy a prompt. This partially explains why Chain-of-Thought (CoT) and structured prompting often outperform standard n-shot prompts. For example, Kartchner et al. [111] performed an error analysis on zero-shot information extraction with GPT, noting that models frequently included extraneous information or generated content outside the task's scope. Similarly, Chervenak [188] observed that LLMs hallucinated roughly 6% of the content and failed to provide references for their claims.

  Models may also refuse to respond or produce vague and unhelpful answers. Giannakopoulos et al. [187] reported that ChatGPT, Bard, and BingChat often provided irrelevant or ambiguous responses. Wilhelm et al. [167] noted frequent error patterns, including confusing diagnoses, vague guidance, and omissions of critical

Table 5: Overview of fine-tuned models used in the selected studies, detailing model architectures, parameter sizes, and whether federated learning was employed, highlighting trends in model selection and scalability within the research landscape.

Study	Fine-tuned model	Parameter size	Federated Learning?
[152]	ALBERT and DistilBERT	12M and 66M	-
[118]	TinyBERT	14.5M	-
[113]	LED	41M	-
[104]	LED	41M	-
[166]	BART	140M	-
[151]	BioClinicalBERT	110M	-
[204]	BERT	110M	✓
[147]	xlm-roberta-base	125M	-
[110]	GLM-large and Chat-GLM	335M and 6B	-
[159]	Alpaca7B	7B	-
[128]	BERT and LLama-7B	110M and 7B	✓
[146]	BioGPT	1.5B	-
[136]	GPT-2	1.5B	-
[137]	GPT-2	1.5B	-
[150]	Phe-BCBERT, Phe-GPT	110M and 6B, 175B (GPT-3)	-
[120]	GPT-3 Ada	175B	-
[160]	GPT-3.5	175B	-

treatments. In critical care scenarios such as meningitis, Fisch et al. [185] found that LLMs refused to answer questions or provided misleading recommendations in 52% of cases.

Model outputs also often neglect their role as clinical decision-support tools. Studies reported that prompted models sometimes failed to provide information on risks, quality-of-life considerations, or potential patient harm [167, 173, 83, 185]. These issues may be attributed to the non-specialized nature of the models used and highlight the challenge of producing accurate, relevant outputs for federated LLMs.

Data quality further affects model performance. Variations in contributor notes, omissions, or inconsistencies can bias model outputs [123]. Beaulieu-Jones et al. [104] showed that free-form clinical notes contained more relevant signals than structured data alone. In federated learning, clients with higher-quality data may disproportionately influence the aggregated model. Shyr et al. [151] and Alsentzer et al. [105] highlighted similar challenges for rare disease prediction and hemorrhage detection, respectively.

Hallucinations remain a significant concern. Cai et al. [78] reported hallucination rates of 42.4% for ChatGPT-3.5 in ophthalmology tasks. Lee et al. [130] found GPT-4 hallucinated in 62% of image reasoning tasks, while a fine-tuned KARA-CXR model hallucinated 25% of the time. Butler et al. [92] observed 4–7% hallucination rates in summarization tasks. Methods to mitigate hallucinations include rule-based grounding [82] and prompt calibration to reduce variability [114].

Explainability is another key challenge. LLMs generally operate as black boxes, making their reasoning opaque [205]. While some studies have sought to improve interpretability [120, 119, 105, 155, 175], no studies have examined explainability in clinical LLMs under federated learning.

Finally, federated LLMs require governance mechanisms to track contributions, manage participant reputation, and ensure data quality. Liu et al. [194] proposed a system for evaluating participant contributions, which could help mitigate malicious activity, enforce the "right to forget" [206], and ensure high-quality data is used for model aggregation.

The deployment of LLMs in healthcare, particularly under federated learning frameworks, faces a multi-faceted set of challenges. Data accessibility and quality impose limitations on model training and domain adaptation, while privacy and ethical considerations necessitate careful governance. Technical barriers include the substantial computational and communication demands of large-scale models, as well as the stochastic and sometimes unpredictable nature of LLM outputs, which can lead to incomplete, inaccurate, or misleading information. Additional concerns involve model explainability, trustworthiness, and alignment with clinical best practices, all of which are critical for safe and effective decision support. Collectively, these challenges underscore that while LLMs hold great promise for

healthcare applications, their integration requires a cautious, systematic approach that balances technical feasibility, ethical responsibility, and clinical reliability.

4.5 Opportunities

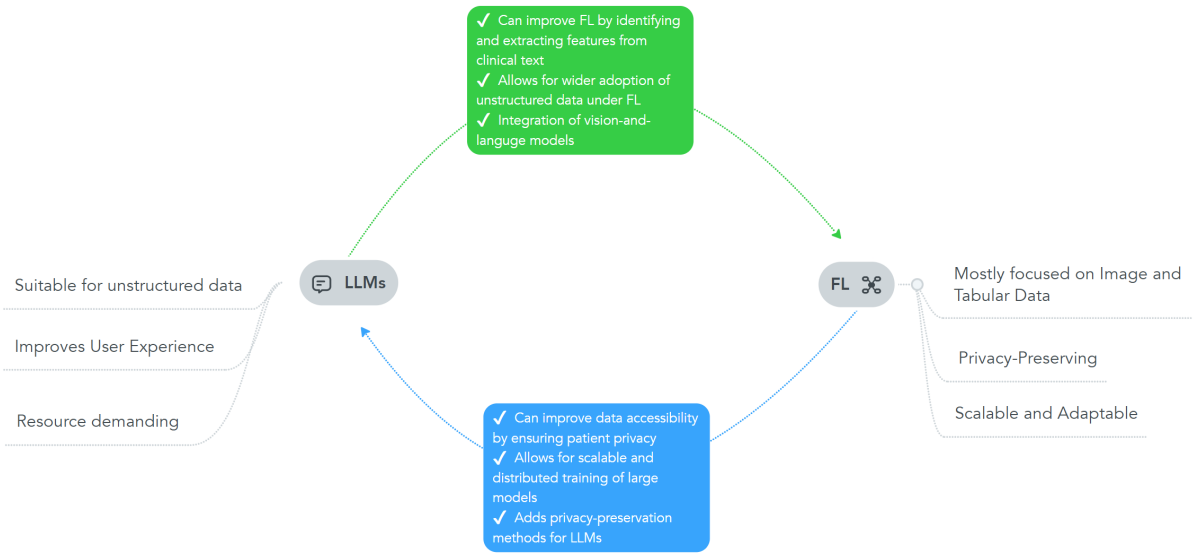

Figure 8: Key advantages of integrating LLMs with FL for healthcare applications include improved data accessibility, enhanced privacy preservation, and broader adoption of unstructured data. The bidirectional relationship indicates that both fields benefit: FL enables scalable, privacy-preserving training for LLMs, while LLMs provide structured text representations for downstream tasks.

Federated learning (FL), as a privacy-preserving technology, has the potential to significantly enhance the capabilities of LLMs in healthcare. In particular, FL enables collaborative and private model training, addressing two critical challenges in healthcare modeling: (1) ensuring data representativeness and sufficient volume, and (2) maintaining confidentiality and privacy of sensitive health information.

Although FL has been extensively explored for tabular and image-based data, text-based EHRs remain largely under-utilized. Of the nearly 1,000 papers reviewed, only 3.2% incorporated language-based data for model training. This section focuses on the potential benefits of applying FL to LLMs, particularly in increasing the availability and quality of text-based biomedical data for modeling.

Traditional model training requires centralizing all data in a single location, which introduces barriers related to privacy, accessibility, and data management. By contrast, FL allows institutions to fine-tune pre-trained language models (PLMs) on task-specific data without sharing the raw data externally. This approach can enhance model reliability, reduce hallucinations, and improve trustworthiness, as models are exposed to domain-relevant information while respecting patient confidentiality.

4.6 Federated LLM training

As observed in our review, most studies relied on pre-trained models, likely due to the substantial costs associated with acquiring training data and running full training loops on edge devices. This creates a clear opportunity for federated LLMs to leverage parameter-efficient fine-tuning (PEFT) techniques. Methods such as LoRA, Instruction Tuning, Prefix Tuning, and P-tuning enable devices with limited computational resources to effectively train LLMs, making decentralized training in an FL setup more feasible.

Table 5 illustrates that most fine-tuned models in the reviewed studies were relatively small, often BERT-type models with parameters in the millions. While these models are flexible and effective, large auto-regressive generative models, such as the GPT family, often contain billions of parameters and require substantial data for successful training. PEFT

approaches present an opportunity to adopt these larger, high-capacity models in healthcare settings while drastically reducing computational costs [207].

Efficient training with PEFT has already been demonstrated in studies like Shin et al.[128], where BERT-type models were fine-tuned on a smartphone using LoRA, reducing the number of trainable parameters. The end-to-end BERT approach under FL consumed approximately 45% of CPU resources and 800MB per training epoch. By freezing the base BERT model and training only a downstream MLP, the authors achieved comparable performance with significantly lower computational overhead.

Prompt-based fine-tuning further reduces the number of training samples needed for clients with limited data. Tan et al.[103] evaluated multiple LLMs for disease classification from radiology reports, showing that without prompt fine-tuning, accuracy dropped sharply from 0.891 to 0.543 when sample size decreased from 8,500 to 100. With prompt-based fine-tuning, accuracy declined only marginally, from 0.889 to 0.8102. These results highlight how PEFT methods empower healthcare participants in federated learning, enabling effective contributions even when their local datasets are small.

Moreover, as data availability and computational resources are not always homogeneous in FL, flexible training and fine-tuning methods can help ensure fair contributions from all participants. Yun et al. [204] examined a federated setup in which a BERT-type model was trained under two separate scenarios: distributed training with FL and fine-tuning with FL. Rather than treating these approaches independently, they can be distributed across clients based on their capabilities. Stronger participants can handle full training rounds, while less capable clients can focus on fine-tuning, all operating on the same shared model.

Under this paradigm, parameter-efficient multi-task learning becomes feasible with federated LLMs. For example, Shin et al. [128] leveraged pre-trained dialog models to improve fine-tuned models for mental health classification. Federated LLMs allow clients with different types of data to contribute complementary knowledge—some providing structural, type, or content variations, while others contribute high-quality annotated diagnoses and outcomes. This distributed, multi-task approach can yield synergistic results that outperform traditional single-client training.

The concept of task-specific training also opens avenues for multimodal LLMs. Traditionally, training multimodal models requires integrating all data types into a single annotated dataset, which can be challenging due to limited data availability. Lee et al. [130] evaluated GPT-4 on radiology images and found that its performance was inferior to the specialized KARA-CXR model, which was trained on private hospital X-ray data. This underscores the difficulty of fine-tuning models on both image and language data simultaneously.

Federated learning offers a promising solution: Lu et al. [168] demonstrated the training of a vision-and-language model using split X-ray datasets across multiple clients, integrating CNN-based image embeddings with BERT-based text embeddings. The resulting model outperformed other state-of-the-art approaches, illustrating the potential of federated LLMs in healthcare. Furthermore, clients could train only specific components of a model—for example, some training the CNN part and others the language part—allowing the aggregation of multimodal capabilities even when individual datasets are limited in scope or type.

A similar concept was explored by Liu et al. [94], where a multimodal LLM was built for X-ray and CT images alongside medical reports. The approach involved pre-training specialist models separately for images, text, and combined image-text data, which were then integrated through a downstream architecture to construct the multimodal model. Incorporating FL into this framework could enable contributors to participate even if they only possess a subset of the data types.

Multimodality represents a direction where the accumulated knowledge of medical imaging under FL could significantly enhance the capabilities of medical LLMs. For instance, Sun et al. [86] developed an AI dietitian to assess patient meals for calorie and ingredient content using images sourced from Chinese social media. Since lifestyle information is sensitive, applying FL could improve data representativeness while preserving privacy. Consequently, the integration of FL-enabled multimodality in language models is a significant, yet still largely untapped, opportunity in healthcare AI.

4.7 LLM Potential Applications in the FL Domain

Language models and federated learning (FL) can mutually benefit from each other. As discussed in Section 4.2, LLMs excel at tasks such as named entity recognition (NER), summarization, and data augmentation. These capabilities can support FL applications based on free-form text records by structuring unstructured data and enabling smaller models to leverage the extracted information.

For example, Bhate et al. [139] utilized a pre-trained transformer to extract structured information tuples from free-form clinical notes. Similarly, Ni et al. [136] employed a BERT model to extract keywords related to mental health from

long-term chatbot conversations. Such strategies facilitate FL adoption in environments with poorly structured data, allowing participants with varying levels of data stewardship to contribute meaningfully. Since a large portion of healthcare information is stored in clinical notes, integrating LLMs in pre-processing can significantly improve access to this largely untapped resource.

LLMs can also be leveraged to generate new features from existing patient data. As observed in Beaulieu-Jones et al. [104], models that extract signals from clinical notes can serve as feature generators, enriching raw EHR content. These features can then be incorporated into existing architectures or fed back into the LLMs themselves, enhancing the predictive power of FL setups.

Finally, the generative capabilities of LLMs present two additional opportunities for FL. First, focusing on privacy preservation, generative LLMs can de-identify and rephrase EHR content while maintaining semantic integrity. Ghanadian et al. [152] demonstrated this approach using a seeding dataset to generate synthetic content for model training. Although models trained solely on synthetic data performed worse than those trained on real data, combining synthetic and real data yielded the best results. Second, synthetic data can improve client data balance, enhance representativeness, and mitigate biases in FL environments where contributions may be unevenly distributed. While not applied in FL, Li et al. [106] also show the potential of LLMs for data augmentation.

5 Discussion

5.1 Future Research Directions

Language models are a relatively recent development in deep learning and, given their scale, can often be trained by institutions with dedicated computational power, like Big Tech companies and research institutions. This makes training an LLM from the ground up an exclusive task and leverages pre-trained models as tools for exploring these models' capabilities in medicine and healthcare.

This trend is backed up by the sheer number of papers that have utilized popular pre-trained models, like ChatGPT, to answer health-related inquiries like patient questions, standardized tests, and patient vignettes. While this is a worthy exploration of the capabilities of LLMs in healthcare, these models are not necessarily trained on biomedical literature, and the knowledge embedded in the models may not reflect important relationships to ensure quality care advice. Thus, these studies illustrate the potential of LLMs to address health-related questions when properly designed for this challenging scenario.

To enable this, there is a demand for methods that improve the ease of fine-tuning or training LLMs on local data. In this sense, we believe a future research direction involves developing resource-efficient and privacy-preserving training or fine-tuning language models on local data under FL. New methods like federated instruction tuning (FedIT)[208], FedPrompt [209], and FedBPT [210] are promising studies that allow taking efficient SOTA models and generating specialized models based on local knowledge while maintaining the participant's privacy.

However, rich quantities of biomedical literature are often restricted to hospitals or research groups and, even then, subjected to strong access control. We envision FL as a foundation technology to enable the construction of collaborative language models trained on high-quality data from millions or billions of patient interactions. These models will be trained on a wealth of data that is not publicly available while still being privacy-preserving by design. These attributes of FL are what we believe to be the antidote to Fries' dataset debt, opening up the possibilities of (very) LLMs.

Of course, even with the promise of "privacy by design" brought in by FL, there is concern about what type of information can be inferred from the models' transmitted data. Under FL, attacks like membership inference attacks, model inversion, or prompt-based data extraction attacks can allow malicious or eavesdropping parties to obtain knowledge of the participating parties and jeopardize the privacy benefits of federating LLMs.

To address this, a future research direction is the development of visualization methods for the aggregated models and weights, allowing for identifying possible knowledge patterns that are embedded. Chen et al.'s [211] results on fair FL can suggest the possibility of identifying parameter patterns in FL that hint at common shared knowledge of federated model contributions, something we believe can enable more improved core knowledge distillation and also allow for de-identification methods on FL contributions.

Likewise, there is also a need to explore encryption and other security measures with LLM. Technologies like Pallier Homomorphic Encryption (PHE) [192], additive homomorphic encryption [195], differential privacy [212], and multipart computing [51] have yet to be explored with Federated LLMs. Little is known about how introducing these security measures can influence contextual embeddings under FL, nor studies about the trade-off between performance and security in this setting.

Even with the challenges presented in this paper, pioneers are already exploring the training of LLMs under the FL paradigm. Yun et al.[204] implemented an FL setup in which a BERT model was fine-tuned to provide a binary classifier based on health records of over 8.000 patients divided among eight separate clients and yielded results akin to the centralized version of the model with reasonable training time. Shin et al.[128] further explored privacy-sensitive geolocation data to provide additional knowledge when fine-tuning a BERT model to identify mental health status. This application showcases one of the potential benefits of introducing Federated LLMs because, as privacy is sought after by design, other types of data can be integrated with language data used by LLMs.

This concept of multimodality is something in which FL methods will become a research trend, and we project that other types of data will be enabled to be integrated with either pre-trained models through downstream models or ground-up multimodal LLMs. Given the experience acquired by many FL studies focusing on image-based methods, time-series-based methods, and other non-language health data, we consider that this cultivated knowledge will soon be transferred to language models under FL. Studies like Lu et al.[168] and Chen et al.[122], in which complex language-vision models are constructed in a private and collaborative manner, show that this is not only feasible but is already being implemented by researchers. It should be a matter of time until their widespread adoption.

Industry-adopted libraries have already begun to provide means of training FedLLMs. As of the second half of 2024, Nvidia's NVFlare has already enabled scalable FL with LLMs [213], as well as the popular FL framework Flower [214]. As frameworks facilitate the implementation of the setups necessary for training and validating Federated LLMs, adoption by developers and practitioners is likely to become more frequent. Moreover, recent developments like DeepSeek[215] show that reducing the resource requirements for training LLMs is feasible and may enable high-performance LLMs in the following years. The China-based group introduced dense distillation of the larger DeepSeek model that may be suitable for deploying LLMs to mobile and edge devices more easily [216].

In addition to the research paths discussed, the personalization of FLLMs represents a critical opportunity, as client heterogeneity is a significant challenge in healthcare settings. Future techniques could enable the adaptation of shared LLMs to the specific data distributions of individual hospitals or departments without sacrificing privacy. Future research should also consider integrating continual learning paradigms into FLLMs. Medical knowledge is not static; systems capable of incremental updates without catastrophic forgetting will be crucial for sustainable deployment. Another venue relies on developing lightweight FLLMs optimized for edge devices, which warrants investigation. Resource-constrained environments such as rural clinics could benefit from techniques like model pruning, quantization, and low-rank adaptation (e.g., LoRA), making it feasible to deploy personalized language models at the point of care.

Furthermore, robustness against adversarial attacks remains an open concern; although privacy risks have been discussed, future research should address poisoning and backdoor vulnerabilities specific to the FLLMs. Finally, explainability and interpretability in FLLMs deserve greater attention. Given the critical nature of clinical decision support, models should provide human-interpretable rationales for their outputs, following emerging trends in explainable AI for healthcare.

Overall, our analysis suggests that technological, data, and application axes are interdependent. Advances in federated LLMs, for example, simultaneously address computational efficiency, data privacy, and model personalization, reflecting how overcoming challenges along one axis can create opportunities along another. These trends make the intersection between FL and language models a fertile research field in the coming years. Researchers will have the potential to enable applications that tap into the relatively unexplored wealth of information present in clinical notes and train collaborative models that leverage the scale of private data. Also, addressing these directions will be key to ensuring that Federated LLMs are effective and private, trustworthy, adaptive, and broadly deployable in real-world healthcare systems.

5.2 Limitations

In systematic reviews, addressing limitations is crucial to ensure readers understand the research's constraints. In our study, several limitations impact the scope and interpretation of findings, which we discuss here to contextualize the results and provide a foundation for further research improvements.

Firstly, our systematic review was bound by a specific research window that covers publications from 2018 to May 2024. This temporal limitation is particularly relevant because it excludes any studies published from June to November 2024, which could introduce recent advancements or data that may alter or expand upon the current findings. The selected time frame allowed us to analyze recent developments thoroughly, yet it inherently limits the study's temporal breadth. Future reviews that include a more current dataset may present additional perspectives or reveal emerging trends absent here.

Secondly, this review is limited by the venues selected for paper inclusion. We focused on a defined set of reputable publication sources to maintain quality and rigor, but this also narrows the range of research studies analyzed. Systematic

reviews often balance scope and depth, and here, we opted for quality and thematic alignment with our research questions. However, this approach excludes potentially relevant studies from less prominent venues or interdisciplinary sources that might offer relevant insights. As such, the results reflect only a segment of the broader research panorama, and additional findings might emerge with a wider venue selection.

The inherent subjectivity of GT coding further contributes to this study's limitations. GT, as a methodology, requires subjective encoding of themes, where researchers must interpret and categorize data in a methodologically consistent way that is reflective of the study's objectives. Different researchers might approach the data from varying perspectives, which could lead to alternative thematic focuses or interpretations. Although we adhere to the guidelines of applying GT to systematic coding protocols as suggested by Bowers et al.[76], the process remains inherently interpretive, which might influence the reproducibility of specific findings and discussion points presented in this paper.

Finally, the synthesis analysis used in this systematic review is well-suited for qualitative data synthesis, but it introduces a potential for interpretive bias. Given that qualitative synthesis involves subjective assessments of thematic patterns, there is a risk that certain biases could influence the results [217]. This method requires the researcher to weigh various aspects of the data, which may lead to an inadvertent emphasis on particular themes or interpretations. While measures were taken to ensure rigor and consistency, readers should interpret these findings with an understanding of the limitations inherent in qualitative synthesis. Future studies may benefit from incorporating complementary methodologies to cross-validate findings and further mitigate interpretive bias.

6 Conclusions

Recent advancements in Federated Learning (FL) and Large Language Models (LLMs) have unlocked unprecedented opportunities across domains, particularly e-health. However, despite the momentum behind both FL and LLMs, there has been a critical gap in the literature of comprehensive review studies focusing on the convergence of FL and LLMs (FLLMs) within this field. This paper addresses this gap by providing the first systematic review of FLLMs in healthcare, using grounded theory to analyze applications, limitations, and future opportunities. By mapping the landscape of current implementations and highlighting underexplored areas such as data modality gaps, model usability, and privacy-preserving techniques, our study lays the foundation for a new research frontier. It paves the way for future work by identifying key enablers and barriers, proposing design considerations for secure, ethical, and effective deployment of FLLMs, and emphasizing the need for interdisciplinary collaboration to realize their transformative potential. Moreover, we identify several promising use cases and architectural trends that point toward the feasibility of scalable, decentralized, and privacy-preserving healthcare systems.

References

[1] Nicola Luigi Bragazzi, Haijiang Dai, Giovanni Damiani, Masoud Behzadifar, Mariano Martini, and Jianhong Wu. How big data and artificial intelligence can help better manage the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, 17(9):3176, May 2020.

[2] Karim Abouelmehdi, Abderrahim Beni-Hessane, and Hayat Khaloufi. Big healthcare data: preserving security and privacy. *Journal of Big Data*, 5(1), January 2018.

[3] Javad Pool, Saeed Akhlaghpour, Farhad Fatehi, and Andrew Burton-Jones. A systematic analysis of failures in protecting personal health data: A scoping review. *International Journal of Information Management*, 74:102719, February 2024.

[4] Jinhyung Lee, Hyeyeong Kim, and Sung J Choi. Do hospital data breaches affect health information technology investment? *DIGITAL HEALTH*, 10, January 2024.

[5] Aviv Segev. Integrating computer vision with web-based knowledge for medical diagnostic assistance. *Expert Systems*, 27(4):247–258, August 2010.

[6] Yi Zhang, Steven W. Su, Branko G. Celler, and Hung T. Nguyen. *Machine Learning-based Nonlinear Model Predictive Control for Heart Rate Response to Exercise*, page 271–285. IMPERIAL COLLEGE PRESS, July 2012.

[7] H. Xu, S. P Stenner, S. Doan, K. B Johnson, L. R Waitman, and J. C Denny. Medex: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1):19–24, January 2010.

[8] Arunim Garg and Vijay Mago. Role of machine learning in medical research: A survey. *Computer Science Review*, 40:100370, 2021.

[9] Kornelia Batko and Andrzej Ślęzak. The use of big data analytics in healthcare. *Journal of big Data*, 9(1):3, 2022.

[10] Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. *Informatics*, 11(3), 2024.

[11] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Publisher correction: Large language models encode clinical knowledge. *Nature*, 620(7973):E19–E19, July 2023.

[12] Gautam Sarma, Hrishikesh Kashyap, and Partha Pratim Medhi. Chatgpt in head and neck oncology-opportunities and challenges. *Indian Journal of Otolaryngology and Head & Neck Surgery*, 76(1):1425–1429, August 2023.

[13] Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. Evaluating the performance of chatgpt in ophthalmology. *Ophthalmology Science*, 3(4):100324, December 2023.

[14] Dhir Gala and Amgad N. Makaryus. The utility of language models in cardiology: A narrative review of the benefits and concerns of chatgpt-4. *International Journal of Environmental Research and Public Health*, 20(15):6438, July 2023.

[15] Maciej Rosoł, Jakub S Gąsior, Jonasz Łaba, Kacper Korzeniewski, and Marcel Młyńczak. Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. *Scientific Reports*, 13(1):20512, 2023.

[16] Soshi Takagi, Takashi Watari, Ayano Erabi, and Kota Sakaguchi. Performance of gpt-3.5 and gpt-4 on the japanese medical licensing examination: Comparison study. *JMIR Medical Education*, 9:e48002, June 2023.

[17] Geoff Currie, Stephanie Robbie, and Peter Tually. Chatgpt and patient information in nuclear medicine: Gpt-3.5 versus gpt-4. *Journal of Nuclear Medicine Technology*, 51(4):307–313, September 2023.

[18] Ryan C. King, Jamil S. Samaan, Yee Hui Yeo, Behram Mody, Dawn M. Lombardo, and Roxana Ghashghaei. Appropriateness of chatgpt in answering heart failure related questions. *Heart, Lung and Circulation*, May 2024.

[19] Abi Aryan, Aakash Kumar Nain, Andrew McMahon, Lucas Augusto Meyer, and Harpreet Singh Sahota. The costly dilemma: Generalization, evaluation and cost-optimal deployment of large language models, 2023.

[20] Jason Fries, Natasha Seelam, Gabriel Altay, Leon Weber, Myungsun Kang, Debajyoti Datta, Ruisi Su, Samuele Garda, Bo Wang, Simon Ott, Matthias Samwald, and Wojciech Kusa. Dataset debt in biomedical language modeling. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*. Association for Computational Linguistics, 2022.

[21] Zongda Wu, Shaolong Xuan, Jian Xie, Chongze Lin, and Chenglang Lu. How to ensure the confidentiality of electronic medical records on the cloud: A technical perspective. *Computers in Biology and Medicine*, 147:105726, August 2022.

[22] Ahmed H. Almulihi, Fawaz Alassery, Asif Irshad Khan, Sarita Shukla, Bineet Kumar Gupta, and Rajeev Kumar. Analyzing the implications of healthcare data breaches through computational technique. *Intelligent Automation & Soft Computing*, 32(3):1763–1779, 2022.

[23] Paige Nong, Julia Adler-Milstein, Sharon Kardia, and Jodyn Platt. Public perspectives on the use of different data types for prediction in healthcare. *Journal of the American Medical Informatics Association*, 31(4):893–900, February 2024.

[24] Jip WTM de Kok, Miguel Á Armengol de la Hoz, Ymke de Jong, Véronique Brokke, Paul WG Elbers, Patrick Thoral, Alejandro Castillejo, Tomás Trenor, Jose M Castellano, Alberto E Bronchalo, et al. A guide to sharing open healthcare data under the general data protection regulation. *Scientific data*, 10(1):404, 2023.

[25] Stanley C. Ahalt, Christopher G. Chute, Karamarie Fecho, Gustavo Glusman, Jennifer Hadlock, Casey Overby Taylor, Emily R. Pfaff, Peter N. Robinson, Harold Solbrig, Casey Ta, Nicholas Tatonetti, and Chunhua Weng. Clinical data: Sources and types, regulatory constraints, applications. *Clinical and Translational Science*, 12(4):329–333, May 2019.

[26] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2016.

[27] Jakub Konečný, Brendan McMahan, and Daniel Ramage. Federated optimization:distributed optimization beyond the datacenter, 2015.

[28] Morgan Ekmefjord, Addi Ait-Mlouk, Sadi Alawadi, Mattias Åkesson, Prashant Singh, Ola Spjuth, Salman Toor, and Andreas Hellander. Scalable federated machine learning with fedn. In *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pages 555–564. IEEE, 2022.

[29] Rodolfo Stoffel Antunes, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. Federated learning for healthcare: Systematic review and architecture proposal. *ACM Trans. Intell. Syst. Technol.*, 13(4), may 2022.

[30] Hao Li, Chengcheng Li, Jian Wang, Aimin Yang, Zezhong Ma, Zunqian Zhang, and Dianbo Hua. Review on security of federated learning and its application in healthcare. *Future Generation Computer Systems*, 144:271–290, July 2023.

[31] Wonsuk Oh and Girish N. Nadkarni. Federated learning in health care using structured medical data. *Advances in Kidney Disease and Health*, 30(1):4–16, January 2023.

[32] Ylona Chun Tie, Melanie Birks, and Karen Francis. Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, 7, January 2019.

[33] Chaochao Chen, Xiaohua Feng, Yuyuan Li, Lingjuan Lyu, Jun Zhou, Xiaolin Zheng, and Jianwei Yin. Integration of large language models and federated learning. *Patterns*, 5(12):101098, December 2024.

[34] Siqi Li, Pinyan Liu, Gustavo G Nascimento, Xinru Wang, Fabio Renato Manzolli Leite, Bibhas Chakraborty, Chuan Hong, Yilin Ning, Feng Xie, Zhen Ling Teo, Daniel Shu Wei Ting, Hamed Haddadi, Marcus Eng Hock Ong, Marco Aurélio Peres, and Nan Liu. Federated and distributed learning applications for electronic health records and structured medical data: a scoping review. *Journal of the American Medical Informatics Association*, 30(12):2041–2049, August 2023.

[35] Dinh C. Nguyen, Quoc-Viet Pham, Pubudu N. Pathirana, Ming Ding, Aruna Seneviratne, Zihuai Lin, Octavia A. Dobre, and Won-Joo Hwang. Federated learning for smart healthcare: A survey, 2021.

[36] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics, 2020.

[37] Nikita Neveditsin, Pawan Lingras, and Vijay Mago. Clinical insights: A comprehensive review of language models in medicine, 2024.

[38] Tongyue Shi, Jun Ma, Zihan Yu, Haowei Xu, Minqi Xiong, Meirong Xiao, Yilin Li, Huiying Zhao, and Guilan Kong. Stochastic parrots or icu experts? large language models in critical care medicine: A scoping review, 2024.

[39] Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang, Themistocles L. Assimes, Libby Hemphill, and Siyuan Ma. A scoping review of using large language models (llms) to investigate electronic health records (ehrs), 2024.

[40] Huizi Yu, Lizhou Fan, Lingyao Li, Jiayan Zhou, Zihui Ma, Lu Xian, Wenyue Hua, Sijia He, Mingyu Jin, Yongfeng Zhang, Ashvin Gandhi, and Xin Ma. Large language models in biomedical and health informatics: A review with bibliometric analysis, 2024.

[41] Maria Helena da Fonseca, Fanny Kovaleski, Claudia Tania Picinin, Bruno Pedroso, and Priscila Rubbo. E-health practices and technologies: A systematic review from 2014 to 2019. *Healthcare*, 9(9):1192, September 2021.

[42] Matija Kovačić, Maja Mutavdžija, and Krešimir Buntak. e-health application, implementation and challenges: A literature review. *Business Systems Research Journal*, 13(1):1–18, June 2022.

[43] Houda Fakhkhari, Bouchaib Bounabat, and Ismail Kassou. Digital health taxonomy. In *Proceedings of the 6th International Conference on Networking, Intelligent Systems & Security*, NISS 2023. ACM, May 2023.

[44] Abdul Khalique Shaikh, Saadat M Alhashmi, Nadia Khalique, Ahmed M. Khedr, Kaamran Raahemifar, and Sadaf Bukhari. Bibliometric analysis on the adoption of artificial intelligence applications in the e-health sector. *DIGITAL HEALTH*, 9:205520762211492, January 2023.

[45] Majeda A Al-Ruzzieh, Omar Ayaad, and Bayan Qaddumi. The role of e-health in improving control and management of covid 19 outbreak: current perspectives. *International Journal of Adolescent Medicine and Health*, 34(4):139–145, August 2020.

[46] Daniel Furtner, Salil Prakash Shinde, Manmohan Singh, Chew Hooi Wong, and Sajita Setia. Digital transformation in medical affairs sparked by the pandemic: Insights and learnings from covid-19 era and beyond. *Pharmaceutical Medicine*, 36(1):1–10, December 2021.

[47] Şölen Zengin and Emel Yontar. Reflections of digital transformation in the health sector in the covid 19 pandemic with the effect of industry 4.0. *Tarsus Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 3(2):136–152, 2022.

[48] Remya Sivan and Zuriati Ahmad Zukarnain. Security and privacy in cloud-based e-health system. *Symmetry*, 13(5):742, April 2021.

[49] Anu Jokinen, Minna Stolt, and Riitta Suhonen. Ethical issues related to ehealth: an integrative review. *Nursing ethics*, 28(2):253–271, 2021.

[50] Alissa Brauneck, Louisa Schmalhorst, Mohammad Mahdi Kazemi Majdabadi, Mohammad Bakhtiari, Uwe Völker, Jan Baumbach, Linda Baumbach, and Gabriele Buchholtz. Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: scoping review. *Journal of Medical Internet Research*, 25:e41588, 2023.

[51] Hiroki Kaminaga, Feras M. Awaysheh, Sadi Alawadi, and Liina Kamm. Mpcfl: Towards multi-party computation for secure federated learning aggregation. In *Proceedings of the IEEE/ACM 16th International Conference on Utility and Cloud Computing*, UCC '23. ACM, December 2023.

[52] Juncen Zhu, Jiannong Cao, Divya Saxena, Shan Jiang, and Houda Ferradi. Blockchain-empowered federated learning: Challenges, solutions, and future directions. *ACM Computing Surveys*, 55(11):1–31, February 2023.

[53] Haokun Fang and Quan Qian. Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet*, 13(4):94, 2021.

[54] Jiajun Wu, Fan Dong, Henry Leung, Zhuangdi Zhu, Jiayu Zhou, and Steve Drew. Topology-aware federated learning in edge computing: A comprehensive survey. *ACM Computing Surveys*, 56(10):1–41, June 2024.

[55] Ali Forootani and Raffaele Iervolino. Asynchronous federated learning: A scalable approach for decentralized machine learning. 2024.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[57] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[58] G. Bharathi Mohan, R. Prasanna Kumar, P. Vishal Krishh, A. Keerthinathan, G. Lavanya, Meka Kavya Uma Meghana, Sheba Sulthana, and Srinath Doss. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems*, 66(9):5047–5070, May 2024.

[59] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[60] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[61] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021.

[62] Silvia Milano, Joshua A McGrane, and Sabina Leonelli. Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5(4):333–334, 2023.

[63] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[64] Yang Liu and Donghai Bi. Quantitative risk analysis of treatment plans for patients with tumor by mining historical similar patients from electronic health records using federated learning. *Risk Analysis*, 43(12):2422–2449, March 2023.

[65] Vinay Pursnani, Yusuf Sermet, Musa Kurt, and Ibrahim Demir. Performance of chatgpt on the us fundamentals of engineering exam: Comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. *Computers and Education: Artificial Intelligence*, 5:100183, 2023.

[66] Akshay Rajaram, Nimesh Patel, Zachary Hickey, Brent Wolfrom, and Joseph Newbigging. Perspectives of undergraduate and graduate medical trainees on documenting clinical notes: Implications for medical education and informatics. *Health Informatics Journal*, 28(2):146045822210934, January 2022.

[67] Charlotte Blease, Leonor Fernandez, Sigall K Bell, Tom Delbanco, and Catherine DesRoches. Empowering patients and reducing inequities: is there potential in sharing clinical notes? *BMJ Quality & Safety*, 29(10):1.8–2, March 2020.

[68] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[69] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019.

[70] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, 2019.

[71] Joschka Haltaufderheide and Robert Ranisch. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *npj Digital Medicine*, 7(1), July 2024.

[72] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021.

[73] Salah Boussen, Jean-Baptiste Denis, Pierre Simeone, David Lagier, Nicolas Bruder, and Lionel Velly. Chatgpt and the stochastic parrot: artificial intelligence in medical research. *British Journal of Anaesthesia*, 131(4):e120–e121, October 2023.

[74] Mauro Giuffrè, Kisung You, and Dennis L Shung. Evaluating chatgpt in medical contexts: the imperative to guard against hallucinations and partial accuracies. *Clinical Gastroenterology and Hepatology*, 22(5):1145–1146, 2024.

[75] Angela Carrera-Rivera, William Ochoa, Felix Larrinaga, and Ganix Lasa. How-to conduct a systematic literature review: A quick guide for computer science research. *MethodsX*, 9:101895, 2022.

[76] Alison W. Bowers and Elizabeth G. Creamer. Core principles of grounded theory in a systematic review of environmental education for secondary students. *International Journal of Social Research Methodology*, 24(6):713–726, September 2020.

[77] Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, and Emma Pierson. Topics, authors, and institutions in large language model research: Trends from 17k arxiv papers, 2023.

[78] Louis Z. Cai, Abdulla Shaheen, Andrew Jin, Riya Fukui, Jonathan S. Yi, Nicolas Yannuzzi, and Chrisfouad Alabiad. Performance of generative large language models on ophthalmology board–style questions. *American Journal of Ophthalmology*, 254:141–149, October 2023.

[79] Marc Cicero Schubert, Wolfgang Wick, and Varun Venkataramani. Performance of large language models on a neurology board–style examination. *JAMA Network Open*, 6(12):e2346721, December 2023.

[80] KeXin Wang, Shengyue Yao, Ziyi Wu, Fei-Yue Wang, Yilun Lin, and Yan Chen. Boosting intelligent diagnostic process in internet hospital: A conversational-ai-enhanced framework. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, October 2023.

[81] Nikhil Gopalakrishnan, Aishwarya Joshi, Jay Chhablani, Naresh Kumar Yadav, Nikitha Gurram Reddy, Padmaja Kumari Rani, Ram Snehith Pulipaka, Rohit Shetty, Shivani Sinha, Vishma Prabhu, and Ramesh Venkatesh. Recommendations for initial diabetic retinopathy screening of diabetic patients using large language model-based artificial intelligence in real-life case scenarios. *International Journal of Retina and Vitreous*, 10(1), January 2024.

[82] Dimitrios P. Panagoulias, Filippos A. Palamidas, Maria Virvou, and George A. Tsihrintzis. Rule-augmented artificial intelligence-empowered systems for medical diagnosis using large language models. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, November 2023.

[83] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *AMIA Annu. Symp. Proc.*, 2023:1105–1114, 2023.

[84] Eunkyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, page 1–16. ACM, April 2023.

[85] Sue Lim and Ralf Schmälzle. Artificial intelligence for health message generation: an empirical study using a large language model (llm) and prompt engineering. *Frontiers in Communication*, 8, May 2023.

[86] Haonan Sun, Kai Zhang, Wei Lan, Qiufeng Gu, Guangxiang Jiang, Xue Yang, Wanli Qin, and Dongran Han. An ai dietitian for type 2 diabetes mellitus management based on large language and image recognition models: Preclinical concept validation study. *Journal of Medical Internet Research*, 25:e51300, November 2023.

[87] Reza Kianian, Deyu Sun, Eric L. Crowell, and Edmund Tsui. The use of large language models to generate education materials about uveitis. *Ophthalmology Retina*, 8(2):195–201, February 2024.

[88] Haifeng Song, Yi Xia, Zhichao Luo, Hui Liu, Yan Song, Xue Zeng, Tianjie Li, Guangxin Zhong, Jianxing Li, Ming Chen, Guangyuan Zhang, and Bo Xiao. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. *Journal of Medical Systems*, 47(1), November 2023.

[89] Eoghan T. Hurley, Bryan S. Crook, Samuel G. Lorentz, Richard M. Danilkowicz, Brian C. Lau, Dean C. Taylor, Jonathan F. Dickens, Oke Anakwenze, and Christopher S. Klifto. Evaluation high-quality of information from chatgpt (artificial intelligence—large language model) artificial intelligence on shoulder stabilization surgery. *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, 40(3):726–731.e6, March 2024.

[90] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. Working with ai to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–29, April 2023.

[91] Nathan P Davies, Robert Wilson, Madeleine S Winder, Simon J Tunster, Kathryn McVicar, Shivan Thakrar, Joe Williams, and Allan Reid. Chatgpt sits the dfph exam: large language model performance and potential to support public health learning. *BMC Medical Education*, 24(1), January 2024.

[92] James J. Butler, Michael C. Harrington, Yixuan Tong, Andrew J. Rosenbaum, Alan P. Samsonov, Raymond J. Walls, and John G. Kennedy. From jargon to clarity: Improving the readability of foot and ankle radiology reports with an artificial intelligence large language model. *Foot and Ankle Surgery*, 30(4):331–337, June 2024.

[93] Hannah Decker, Karen Trang, Joel Ramirez, Alexis Colley, Logan Pierce, Melissa Coleman, Tasce Bongiovanni, Genevieve B. Melton, and Elizabeth Wick. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Network Open*, 6(10):e2336997, October 2023.

[94] Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, Yang Yang, Lei Clifton, and David A. Clifton. A medical multimodal large language model for future pandemics. *npj Digital Medicine*, 6(1), December 2023.

[95] Panagiotis Tsoutsanis and Aristotelis Tsoutsanis. Evaluation of large language model performance on the multi-specialty recruitment assessment (msra) exam. *Computers in Biology and Medicine*, 168:107794, January 2024.

[96] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, July 2023.

[97] Jad Abi-Rafeh, Vanessa J. Mroueh, Brian Bassiri-Tehrani, Jacob Marks, Roy Kazan, and Foad Nahai. Complications following body contouring: Performance validation of bard, a novel ai large language model, in triaging and managing postoperative patient concerns. *Aesthetic Plastic Surgery*, 48(5):953–976, January 2024.

[98] Jonathan Kottlors, Grischa Bratke, Philip Rauen, Christoph Kabbasch, Thorsten Persigehl, Marc Schlamann, and Simon Lennartz. Feasibility of differential diagnosis based on imaging patterns using a large language model. *Radiology*, 308(1), July 2023.

[99] Shunsuke Koga, Nicholas B. Martin, and Dennis W. Dickson. Evaluating the performance of large language models: Chatgpt and google bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. *Brain Pathology*, 34(3), August 2023.

[100] Manuela Benary, Xing David Wang, Max Schmidt, Dominik Soll, Georg Hilfenhaus, Mani Nassir, Christian Sigler, Maren Knödler, Ulrich Keller, Dieter Beule, Ulrich Keilholz, Ulf Leser, and Damian T. Rieke. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11):e2343689, November 2023.

[101] Liang Zhang, Syoichi Tashiro, Masahiko Mukaino, and Shin Yamada. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. *Journal of Rehabilitation Medicine*, 55:jrm13373, September 2023.

[102] Ivan Civettini, Arianna Zappaterra, Bianca Maria Granelli, Giovanni Rindone, Andrea Aroldi, Stefano Bonfanti, Federica Colombo, Marilena Fedele, Giovanni Grillo, Matteo Parma, Paola Perfetti, Elisabetta Terruzzi, Carlo Gambacorti-Passerini, Daniele Ramazzotti, and Fabrizio Cavalca. Evaluating the performance of large language models in haematopoietic stem cell transplantation decision-making. *British Journal of Haematology*, 204(4):1523–1528, December 2023.

[103] Ryan Shea Ying Cong Tan, Qian Lin, Guat Hwa Low, Ruixi Lin, Tzer Chew Goh, Christopher Chu En Chang, Fung Fung Lee, Wei Yin Chan, Wei Chong Tan, Han Jieh Tey, Fun Loon Leong, Hong Qi Tan, Wen Long Nei, Wen Yee Chay, David Wai Meng Tai, Gillianne Geet Yi Lai, Lionel Tim-Ee Cheng, Fuh Yong Wong, Matthew Chin Heng Chua, Melvin Lee Kiang Chua, Daniel Shao Weng Tan, Choon Hua Thng, Iain Bee Huat Tan, and Hwee Tou Ng. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. *Journal of the American Medical Informatics Association*, 30(10):1657–1664, July 2023.

[104] Brett K Beaulieu-Jones, Mauricio F Villamar, Phil Scordis, Ana Paula Bartmann, Waqar Ali, Benjamin D Wissel, Emily Alsentzer, Johann de Jong, Arijit Patra, and Isaac Kohane. Predicting seizure recurrence after an initial seizure-like episode from routine clinical notes using large language models: a retrospective cohort study. *The Lancet Digital Health*, 5(12):e882–e894, December 2023.

[105] Emily Alsentzer, Matthew J. Rasmussen, Romy Fontoura, Alexis L. Cull, Brett Beaulieu-Jones, Kathryn J. Gray, David W. Bates, and Vesela P. Kovacheva. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *npj Digital Medicine*, 6(1), November 2023.

[106] Rumeng Li, Xun Wang, and Hong Yu. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023.

[107] Yiqi Wang, Jinmei Zuo, Chao Duan, Hao Peng, Jia Huang, Liang Zhao, Li Zhang, and Zhiqiang Dong. Large language models assisted multi-effect variants mining on cerebral cavernous malformation familial whole genome sequencing. *Computational and Structural Biotechnology Journal*, 23:843–858, December 2024.

[108] Jeffrey Michael Franc, Lenard Cheng, Alexander Hart, Ryan Hata, and Atilla Hertelendy. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (chatgpt) to perform emergency department triage using the canadian triage and acuity scale. *Canadian Journal of Emergency Medicine*, 26(1):40–46, January 2024.

[109] Jie Yuan, Rui Tang, Xiaoming Jiang, and Xiaoli Hu. Large language models for healthcare data augmentation: An example on patient-trial matching. In *AMIA Annual Symposium proceedings. AMIA Symposium*, pages 1324–1333. AMIA, January 2023.

[110] WeiLong Wu, Miao Li, Ji Wu, Ming Ni, and Huishu Yuan. Learning to generate radiology findings from impressions based on large language model. In *2023 IEEE International Conference on Big Data (BigData)*, volume 33, page 2550–2554. IEEE, December 2023.

[111] David Kartchner, Selvi Ramalingam, Irfan Al-Hussaini, Olivia Kronick, and Cassie Mitchell. Zero-shot information extraction for clinical meta-analysis using large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Association for Computational Linguistics, 2023.

[112] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G. Nestor, Ali Soroush, Pierre A. Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F. Rousseau, Chunhua Weng, and Yifan Peng. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1), August 2023.

[113] John Giorgi, Augustin Toma, Ronald Xie, Sondra Chen, Kevin An, Grace Zheng, and Bo Wang. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, 2023.

[114] Yu-Neng Chuang, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Spec: A soft prompt-based calibration on performance variability of large language model in clinical notes summarization. *Journal of Biomedical Informatics*, 151:104606, March 2024.

[115] Varun Nair, Elliot Schumacher, and Anitha Kannan. Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, page 200–217. Association for Computational Linguistics, 2023.

[116] Gjorgjina Cenikj, Lidija Strojnik, Risto Angelski, Nives Ogrinc, Barbara Koroušić Seljak, and Tome Eftimov. From language models to large-scale food and biomedical knowledge graphs. *Scientific Reports*, 13(1), May 2023.

[117] Xi Chen, Cheng Li, Ziyuan Wang, Yixin Zhou, and Ming Chu. Computational screening of biomarkers and potential drugs for arthrofibrosis based on combination of sequencing and large nature language model. *Journal of Orthopaedic Translation*, 44:102–113, January 2024.

[118] Vincenzo Moscato, Marco Postiglione, Carlo Sansone, and Giancarlo Sperlí. Taughtnet: Learning multi-task biomedical named entity recognition from single-task teachers. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2512–2523, May 2023.

[119] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

[120] Hirak Mazumdar, Chinmay Chakraborty, MSVPJ Sathvik, sabyasachi Mukhopadhyay, and Prasanta K Panigrahi. Gptfx: A novel gpt-3 based framework for mental health detection and explanations. *IEEE Journal of Biomedical and Health Informatics*, page 1–8, 2024.

[121] Ross O'Hagan, Randie H. Kim, Brian J. Abittan, Stella Caldas, Jonathan Ungar, and Benjamin Ungar. Trends in accuracy and appropriateness of alopecia areata information obtained from a popular online large language model, chatgpt. *Dermatology*, 239(6):952–957, 2023.

[122] Jieying Chen and Rong Pan. Medical report generation based on multimodal federated learning. *Computerized Medical Imaging and Graphics*, 113:102342, April 2024.

[123] Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch. Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*, 112:59–67, April 2018.

[124] Thomas Borger, Pablo Mosteiro, Heysem Kaya, Emil Rijcken, Albert Ali Salah, Floortje Scheepers, and Marco Spruit. Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting. *Expert Systems with Applications*, 199:116720, August 2022.

[125] Olivia Choudhury, Yoonyoung Park, Theodoros Salonidis, Aris Gkoulalas-Divanis, Issa Sylla, and Amar K Das. Predicting adverse drug reactions on distributed health data using federated learning. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2019:313—322, 2019.

[126] Athanasios Sarlas, Alexandros Kalafatelis, Georgios Alexandridis, Michail-Alexandros Kourtis, and Panagiotis Trakadas. Exploring federated learning for speech-based parkinson's disease detection. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, ARES 2023. ACM, August 2023.

[127] Wenqing Wei, Zhengdong Yang, Yuan Gao, Jiyi Li, Chenhui Chu, Shogo Okada, and Sheng Li. Fedcpc: An effective federated contrastive learning method for privacy preserving early-stage alzheimers speech detection. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, December 2023.

[128] Jaemin Shin, Hyungjun Yoon, Seungjoo Lee, Sungjoon Park, Yunxin Liu, Jinho Choi, and Sung-Ju Lee. Fedtherapist: Mental health monitoring with user-generated linguistic expressions on smartphones via federated learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023.

[129] Usman Ahmed, Jerry Chun-Wei Lin, and Gautam Srivastava Srivastava. Hyper-graph attention based federated learning methods for use in mental health detection. *IEEE Journal of Biomedical and Health Informatics*, 27(2):768–777, February 2023.

[130] Kyu Hong Lee, Ro Woon Lee, and Ye Eun Kwon. Validation of a deep learning chest x-ray interpretation model: Integrating large-scale ai and large language models for comparative analysis with chatgpt. *Diagnostics*, 14(1):90, December 2023.

[131] Liang Zhang, Syoichi Tashiro, Masahiko Mukaino, and Shin Yamada. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. *Journal of Rehabilitation Medicine*, 55:jrm13373, September 2023.

[132] Arun James Thirunavukarasu, Refaat Hassan, Shathar Mahmood, Rohan Sanghera, Kara Barzangi, Mohanned El Mukashfi, and Sachin Shah. Trialling a large language model (chatgpt) in general practice with the applied knowledge test: Observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education*, 9:e46599, April 2023.

[133] Vera Sorin, Eyal Klang, Miri Sklair-Levy, Israel Cohen, Douglas B. Zippel, Nora Balint Lahat, Eli Konen, and Yiftach Barash. Large language model (chatgpt) as a support tool for breast tumor board. *npj Breast Cancer*, 9(1), May 2023.

[134] Krithi Pushpanathan, Zhi Wei Lim, Samantha Min Er Yew, David Ziyou Chen, Hazel Anne Hui'En Lin, Jocelyn Hui Lin Goh, Wendy Meihua Wong, Xiaofei Wang, Marcus Chun Jin Tan, Victor Teck Chang Koh, and Yih-Chung Tham. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience*, 26(11):108163, November 2023.

[135] Adi Lahat, Eyal Shachar, Benjamin Avidan, Zina Shatz, Benjamin S. Glicksberg, and Eyal Klang. Evaluating the use of large language model in identifying top research questions in gastroenterology. *Scientific Reports*, 13(1), March 2023.

[136] Yeming Ni, Ruyi Ding, Yuqing Chen, Hanchao Hou, and Shiguang Ni. Focusing on needs: A chatbot-based emotion regulation tool for adolescents. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, volume 1, page 2295–2300. IEEE, October 2023.

[137] Elisa Terumi Rubel Schneider, Joao Vitor Andrioli de Souza, Yohan Bonescki Gumiel, Claudia Moro, and Emerson Cabrera Paraiso. A gpt-2 language model for biomedical texts in portuguese. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, volume 264, page 474–479. IEEE, June 2021.

[138] Zhi Wei Lim, Krithi Pushpanathan, Samantha Min Er Yew, Yien Lai, Chen-Hsin Sun, Janice Sing Harn Lam, David Ziyou Chen, Jocelyn Hui Lin Goh, Marcus Chun Jin Tan, Bin Sheng, Ching-Yu Cheng, Victor Teck Chang Koh, and Yih-Chung Tham. Benchmarking large language models' performances for myopia care: a comparative analysis of chatgpt-3.5, chatgpt-4.0, and google bard. *eBioMedicine*, 95:104770, September 2023.

[139] Neel Jitesh Bhate, Ansh Mittal, Zhe He, and Xiao Luo. Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using gpt model. In *2023 IEEE International Conference on Big Data (BigData)*, volume 2014, page 1476–1480. IEEE, December 2023.

[140] Kefaya Sabaneh, Momen Abu Salameh, Fatima Khaleel, Mohammad M. Herzallah, Joman Y. Natsheh, and Mohammed Maree. Early risk prediction of depression based on social media posts in arabic. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, page 595–602. IEEE, November 2023.

[141] Thomas Labbé, Pierre Castel, Jean-Michel Sanner, and Majd Saleh. Chatgpt for phenotypes extraction: one model to rule them all? In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, volume 2009, page 1–4. IEEE, July 2023.

[142] Branden R. Sosa, Michelle Cung, Vincentius J. Suhardi, Kyle Morse, Andrew Thomson, He S. Yang, Sravisht Iyer, and Matthew B. Greenblatt. Capacity for large language model chatbots to aid in orthopedic management, research, and patient queries. *Journal of Orthopaedic Research*, 42(6):1276–1282, January 2024.

[143] Carrie Ye, Elric Zweck, Zechen Ma, Justin Smith, and Steven Katz. Doctor versus artificial intelligence: Patient and physician evaluation of large language model responses to rheumatology patient questions in a <scp>cross-sectional</scp> study. *Arthritis & Rheumatology*, 76(3):479–484, January 2024.

[144] Surabhi Datta, Kyeryoung Lee, Hunki Paek, Frank J Manion, Nneka Ofoegbu, Jingcheng Du, Ying Li, Liang-Chin Huang, Jingqi Wang, Bin Lin, Hua Xu, and Xiaoyan Wang. Autocriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *Journal of the American Medical Informatics Association*, 31(2):375–385, November 2023.

[145] Tim Reason, William Rawlinson, Julia Langham, Andy Gimblett, Bill Malcolm, and Sven Klijn. Artificial intelligence to automate health economic modelling: A case study to evaluate the potential application of large language models. *PharmacoEconomics - Open*, 8(2):191–203, February 2024.

[146] K M Sajjadul Islam, Ayesha Siddika Nipu, Praveen Madiraju, and Priya Deshpande. Autocompletion of chief complaints in the electronic health records using large language models. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, December 2023.

[147] Simon Meoni, Eric De la Clergerie, and Theo Ryffel. Large language models as instructors: A study on multilingual clinical entity extraction. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, page 178–190. Association for Computational Linguistics, 2023.

[148] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6(1), November 2023.

[149] Ryan Shea Ying Cong Tan, Qian Lin, Guat Hwa Low, Ruixi Lin, Tzer Chew Goh, Christopher Chu En Chang, Fung Fung Lee, Wei Yin Chan, Wei Chong Tan, Han Jieh Tey, Fun Loon Leong, Hong Qi Tan, Wen Long Nei, Wen Yee Chay, David Wai Meng Tai, Gillianne Geet Yi Lai, Lionel Tim-Ee Cheng, Fuh Yong Wong, Matthew Chin Heng Chua, Melvin Lee Kiang Chua, Daniel Shao Weng Tan, Choon Hua Thng, Iain Bee Huat Tan, and

Hwee Tou Ng. Inferring cancer disease response from radiology reports using large language models with data augmentation and prompting. *Journal of the American Medical Informatics Association*, 30(10):1657–1664, July 2023.

[150] Jingye Yang, Cong Liu, Wendy Deng, Da Wu, Chunhua Weng, Yunyun Zhou, and Kai Wang. Enhancing phenotype recognition in clinical notes using large language models: Phenobcbert and phenogpt. *Patterns*, 5(1):100887, January 2024.

[151] Cathy Shyr, Yan Hu, Lisa Bastarache, Alex Cheng, Rizwan Hamid, Paul Harris, and Hua Xu. Identifying and extracting rare diseases and their phenotypes with large language models. *Journal of Healthcare Informatics Research*, 8(2):438–461, January 2024.

[152] Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEE Access*, 12:14350–14363, 2024.

[153] Leonard Ruocco, Yuqian Zhuang, Raymond Ng, Richard J Munthali, Kristen L Hudec, Angel Y Wang, Melissa Vereschagin, and Daniel V Vigo. A platform for connecting social media data to domain-specific topics using large language models: an application to student mental health. *JAMIA Open*, 7(1), January 2024.

[154] Hao Li, Yuping Wu, Viktor Schlegel, Riza Batista-Navarro, Thanh-Tung Nguyen, Abhinav Ramesh Kashyap, Xiaojun Zeng, Daniel Beck, Stefan Winkler, and Goran Nenadic. Pulsar: Pre-training with extracted healthcare terms for summarising patients' problems and data augmentation with black-box large language models, 2023.

[155] Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron Wallace. Chill: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023.

[156] Gloria Wu, Weichen Zhao, Adrial Wong, and David A Lee. Patients with floaters: Answers from virtual assistants and large language models. *DIGITAL HEALTH*, 10, January 2024.

[157] Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. Large language models propagate race-based medicine. *npj Digital Medicine*, 6(1), October 2023.

[158] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

[159] Yixing Jiang, Jeremy A. Irvin, Andrew Y. Ng, and James Zou. Vetllm: Large language model for predicting diagnosis from veterinary notes. In *Biocomputing 2024*, page 120–133. WORLD SCIENTIFIC, December 2023.

[160] Zhiyu Chen, Yujie Lu, and William Wang. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 4295–4304. Association for Computational Linguistics, 2023.

[161] Pritam Mukherjee, Benjamin Hou, Ricardo B. Lanfredi, and Ronald M. Summers. Feasibility of using the privacy-preserving large language model vicuna for labeling radiology reports. *Radiology*, 309(1), October 2023.

[162] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *npj Digital Medicine*, 5(1), December 2022.

[163] Lleayem Nazario-Johnson, Hossam A. Zaki, and Glenn A. Tung. Use of large language models to predict neuroimaging. *Journal of the American College of Radiology*, 20(10):1004–1009, October 2023.

[164] Weipeng Zhou, Majid Afshar, Dmitriy Dligach, Yanjun Gao, and Timothy Miller. Improving the transferability of clinical note section classification models with bert and large language model ensembles. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, page 125–130. Association for Computational Linguistics, 2023.

[165] Hao Li, Yuping Wu, Viktor Schlegel, Riza Batista-Navarro, Thanh-Tung Nguyen, Abhinav Ramesh Kashyap, Xiaojun Zeng, Daniel Beck, Stefan Winkler, and Goran Nenadic. Pulsar: Pre-training with extracted healthcare terms for summarising patients' problems and data augmentation with black-box large language models, 2023.

[166] Ashwyn Sharma, David Feldman, and Aneesh Jain. Team cadence at mediqa-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, page 228–235. Association for Computational Linguistics, 2023.

[167] Theresa Isabelle Wilhelm, Jonas Roos, and Robert Kaczmarczyk. Large language models for therapy recommendations across 3 clinical specialties: Comparative study. *Journal of Medical Internet Research*, 25:e49324, October 2023.

[168] Siyu Lu, Zheng Liu, Tianlin Liu, and Wangchunshu Zhou. Scaling-up medical vision-and-language representation learning with federated learning. *Engineering Applications of Artificial Intelligence*, 126:107037, November 2023.

[169] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[170] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022.

[171] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022.

[172] OpenAI. *OpenAI Documentation: Prompt Engineering Guide*, 2024. Accessed: 2024-10-28.

[173] Isaac A. Bernstein, Youchen (Victor) Zhang, Devendra Govil, Iyad Majid, Robert T. Chang, Yang Sun, Ann Shue, Jonathan C. Chou, Emily Schehlein, Karen L. Christopher, Sylvia L. Groth, Cassie Ludwig, and Sophia Y. Wang. Comparison of ophthalmologist and large language model chatbot responses to online patient eye care questions. *JAMA Network Open*, 6(8):e2330320, August 2023.

[174] Emilie Steerling, Elin Siira, Per Nilsen, Petra Svedberg, and Jens Nygren. Implementing ai in healthcare—the relevance of trust: a scoping review. *Frontiers in Health Services*, 3, August 2023.

[175] Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine*, 7(1), January 2024.

[176] Ethan Schonfeld, Aaradhya Pant, Aaryan Shah, Sina Sadeghzadeh, Dhiraj Pangal, Adrian Rodrigues, Kelly Yoo, Neelan Marianayagam, Ghani Haider, and Anand Veeravagu. Evaluating computer vision, large language, and genome-wide association models in a limited sized patient cohort for pre-operative risk stratification in adult spinal deformity surgery. *Journal of Clinical Medicine*, 13(3):656, January 2024.

[177] Teerapaun Tanprasert and David Kauchak. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, page 1–14. Association for Computational Linguistics, 2021.

[178] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, ACL '02, page 311. Association for Computational Linguistics, 2001.

[179] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[180] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[181] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[182] D. Charnock, S. Shepperd, G. Needham, and R. Gann. Discern: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology & Community Health*, 53(2):105–111, February 1999.

[183] D. Charnock. Learning to discern online: applying an appraisal tool to health websites in a workshop setting. *Health Education Research*, 19(4):440–446, May 2004.

[184] Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Erwin Vu, Galina F. Fischer, Krisztian Süveg, Markus Glatzer, Elena Riggenbach, Hông-Linh Hà, and Nikola Cihoric. Exploring capabilities of large language models such as chatgpt in radiation oncology. *Advances in Radiation Oncology*, 9(3):101400, March 2024.

[185] Urs Fisch, Paulina Kliem, Pascale Grzonka, and Raoul Sutter. Performance of large language models on advocating the management of meningitis: a comparative qualitative study. *BMJ Health & Care Informatics*, 31(1):e100978, February 2024.

[186] Mohita Chowdhury, Ernest Lim, Aisling Higham, Rory McKinnon, Nikoletta Ventoura, Yajie He, and Nick De Pennington. Can large language models safely address patient questions following cataract surgery? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, 2023.

[187] Kostis Giannakopoulos, Argyro Kavadella, Anas Aaqel Salim, Vassilis Stamatopoulos, and Eleftherios G Kaklamanos. Evaluation of the performance of generative ai large language models chatgpt, google bard, and microsoft bing chat in supporting evidence-based dentistry: Comparative mixed methods study. *Journal of Medical Internet Research*, 25:e51580, December 2023.

[188] Joseph Chervenak, Harry Lieman, Miranda Blanco-Breindel, and Sangita Jindal. The promise and peril of using a large language model to obtain clinical information: Chatgpt performs strongly as a fertility counseling tool with limitations. *Fertility and Sterility*, 120(3):575–583, September 2023.

[189] Dat Duong and Benjamin D. Solomon. Analysis of large-language model versus human performance for genetics questions. *European Journal of Human Genetics*, 32(4):466–468, May 2023.

[190] Brendin R. Beaulieu-Jones, Margaret T. Berrigan, Sahaj Shah, Jayson S. Marwaha, Shuo-Lun Lai, and Gabriel A. Brat. Evaluating capabilities of large language models: Performance of gpt-4 on surgical knowledge assessments. *Surgery*, 175(4):936–942, April 2024.

[191] Romain Bey, Romain Goussault, François Grolleau, Mehdi Benchoufi, and Raphaël Porcher. Fold-stratified cross-validation for unbiased and privacy-preserving federated learning. *Journal of the American Medical Informatics Association*, 27(8):1244–1251, July 2020.

[192] Vineetha Pais, Santhosha Rao, Balachandra Muniyal, and Sheng Yun. Fedicu: a federated learning model for reducing the medication prescription errors in intensive care units. *Cogent Engineering*, 11(1), January 2024.

[193] Jafar A. Alzubi, Omar A. Alzubi, Ashish Singh, and Manikandan Ramachandran. Cloud-iiot-based electronic health record privacy-preserving by cnn and blockchain-enabled federated learning. *IEEE Transactions on Industrial Informatics*, 19(1):1080–1087, January 2023.

[194] Zelei Liu, Yuanyuan Chen, Yansong Zhao, Han Yu, Yang Liu, Renyi Bao, Jinpeng Jiang, Zaiqing Nie, Qian Xu, and Qiang Yang. Contribution-aware federated learning for smart healthcare. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12396–12404, June 2022.

[195] Zeyue Xue, Pan Zhou, Zichuan Xu, Xiumin Wang, Yulai Xie, Xiaofeng Ding, and Shiping Wen. A resource-constrained and privacy-preserving edge-computing-enabled clinical decision system: A federated reinforcement learning approach. *IEEE Internet of Things Journal*, 8(11):9122–9138, June 2021.

[196] Jiachun Li, Yan Meng, Lichuan Ma, Suguo Du, Haojin Zhu, Qingqi Pei, and Xuemin Shen. A federated learning based privacy-preserving smart healthcare system. *IEEE Transactions on Industrial Informatics*, 18(3):2021–2031, March 2022.

[197] Soroosh Tayebi Arasteh, Cristian David Ríos-Urrego, Elmar Nöth, Andreas Maier, Seung Hee Yang, Jan Rusz, and Juan Rafael Orozco-Arroyave. Federated Learning for Secure Development of AI Models for Parkinson's Disease Detection Using Speech from Different Languages. In *Proc. INTERSPEECH 2023*, pages 5003–5007, 2023.

[198] Geun Hyeong Lee and Soo-Yong Shin. Federated learning on clinical benchmark data: Performance assessment. *Journal of Medical Internet Research*, 22(10):e20891, October 2020.

[199] Mingyi Li, Xiao Zhang, Haochao Ying, Yang Li, Xu Han, and Dongxiao Yu. Data quality aware hierarchical federated reinforcement learning framework for dynamic treatment regimes. In *2023 IEEE International Conference on Data Mining (ICDM)*, page 1103–1108. IEEE, December 2023.

[200] Weishen Pan, Zhenxing Xu, Suraj Rajendran, and Fei Wang. An adaptive federated learning framework for clinical risk prediction with electronic health records from multiple hospitals. *Patterns*, 5(1):100898, January 2024.

[201] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv, 2024.

[202] Kimia Tuz Zaman, Wordh Ul Hasan, Juan Li, and Cui Tao. Empowering caregivers of alzheimer's disease and related dementias (adrd) with a gpt-powered voice assistant: Leveraging peer insights from social media. In *2023 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, July 2023.

[203] Dimitrios P. Panagoulias, Maria Virvou, and George A. Tsihrintzis. Augmenting large language models with rules for enhanced domain-specific interactions: The case of medical diagnosis. *Electronics*, 13(2):320, January 2024.

[204] Won Joon Yun, Samuel Kim, and Joongheon Kim. Multi-site clinical federated learning using recursive and attentive models and nvflare. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, July 2023.

[205] James E. Dobson. On reading and interpreting black box deep neural networks. *International Journal of Digital Humanities*, 5(2–3):431–449, November 2023.

[206] Leijie Wu, Song Guo, Junxiao Wang, Zicong Hong, Jie Zhang, and Yaohong Ding. Federated unlearning: Guarantee the right of clients to forget. *IEEE Network*, 36(5):129–135, September 2022.

[207] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, March 2023.

[208] Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Yufan Zhou, Guoyin Wang, and Yiran Chen. Towards building the federated gpt: Federated instruction tuning, 2024.

[209] Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. Fedprompt: Communication-efficient and privacy preserving prompt tuning in federated learning, 2023.

[210] Jingwei Sun, Ziyue Xu, Hongxu Yin, Dong Yang, Daguang Xu, Yiran Chen, and Holger R. Roth. Fedbpt: Efficient federated black-box prompt tuning for large language models, 2023.

[211] Yuhang Chen, Wenke Huang, and Mang Ye. Fair federated learning under domain skew with local consistency and domain diversity, 2024.

[212] Yue Cui, Zhuohang Li, Luyang Liu, Jiaxin Zhang, and Jian Liu. Privacy-preserving speech-based depression diagnosis via federated learning. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, July 2022.

[213] NVIDIA. Scalable federated learning with nvidia flare for enhanced llm performance, May 2024.

[214] Flower. Introducing flowerllm.

[215] DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024.

[216] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang

Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[217] Kathy Charmaz and Robert Thornberg. The pursuit of quality in grounded theory. *Qualitative research in psychology*, 18(3):305–327, 2021.