Article published online: 2025-04-08

Natural Language Processing for Digital Health in the Era of Large Language Models

Abeed Sarker[1], Rui Zhang[2], Yanshan Wang[3], Yunyu Xiao[4], Sudeshna Das[1], Dalton Schutte[2], David Oniani[3], Qianqian Xie[5], and Hua Xu[5]

1 Emory University, Atlanta, GA, USA {abeed.sarker,sudeshna.das}@emory.edu
2 University of Minnesota, Minneapolis, MN, USA {zhan1386,schut184}@umn.edu
3 University of Pittsburgh, Pittsburgh, PA, USA {yanshan.wang,davidoniani}@pitt.edu
4 Cornell University, New York, NY, USA {yux4008@}med.cornell.edu
5 Yale University, New Haven, CT, USA {qianqian.xie,hua.xu}@yale.edu

Summary

Objectives: Large language models (LLMs) are revolutionizing the natural language pro-cessing (NLP) landscape within health-care, prompting the need to synthesize the latest ad-vancements and their diverse medical applications. We attempt to summarize the current state of research in this rapidly evolving space.
Methods: We conducted a review of the most recent studies on biomedical NLP facilitated by LLMs, sourcing literature from PubMed, the Association for Computational Linguistics Anthology, IEEE Explore, and Google Scholar (the latter particularly for preprints). Given the ongoing exponential growth in LLM-related publications, our survey was inherently selective. We attempted to abstract key findings in terms of (i) LLMs customized for medical texts, and (ii) the type of medical text being leveraged by LLMs, namely medical literature, electronic health records (EHRs), and social media. In addition to technical details, we touch upon topics such as privacy, bias, interpretability, and equitability.
Results: We observed that while general-purpose LLMs (e.g., GPT-4) are most popular, there is a growing trend in training or customizing open-source LLMs for specific biomedi-cal texts and tasks. Several promising open-source LLMs are currently available, and appli-cations involving EHRs and biomedical literature are more prominent relative to noisier data sources such as social media. For supervised classification and named entity recogni-tion tasks, traditional (encoder only) transformer-based models still outperform new-age LLMs, and the latter are typically suited for few-shot settings and generative tasks such as summarization. There is still a paucity of research on evaluation, bias, privacy, reproduci-bility, and equitability of LLMs.
Conclusions: LLMs have the potential to transform NLP tasks within the broader medical domain. While technical progress continues, biomedical application focused research must prioritize aspects not necessarily related to performance such as task-ori-ented evaluation, bias, and equitable use.

Keywords

Deep Learning; Large Language Models; Natural Language Processing; Medical Informatics.

Yearb Med Inform 2023: http://dx.doi.org/10.1055/s-0044-1800750

1. Introduction

Language models encode characteristics of a language in a machine-readable form—as numeric vectors. Over approximately the last decade, two major advances in language modeling have transformed how they are used in natural language processing (NLP) research and applications:

- the ability to capture word- or phrase-level semantics (meanings) in the form of dense vectors (semantically similar lexical expressions appear close to each other in a vector space) [1]; and, later;
- the ability to capture contextual variations in meanings via transformer-based models [2].

The transition from (1) to (2) was particularly impactful—pretrained transformer models enabled systems to advance the state-of-the-art in many standardized NLP tasks [3,4]. Broadly speaking, current state-of-the-art language models attempt to capture probability distributions over sequences of words in a given language. They do this by learning patterns and relationships between words from large amounts of text-based data. Popular mechanisms for learning such patterns include masked language modeling (e.g., predicting masked words from their contexts), and autoregressive pretraining [5] (e.g., next word prediction). Since such pretraining methods are self-supervised, and text-based data are abundant, it is possible to learn powerful representations (e.g., BERT [2] and GPT-3 [6]). Starting with BERT, the ability to further train existing pretrained transformer models led to the creation of many models—some customized for texts in restricted domains, such as the medical domain [7,8]. Within this period that observed unprecedented advances in language modeling, early advances were primarily in encoder-only models, and more recent advances have been in encoder-decoder or decoder-only (generative) models [9].

Over the last five years, the concept of large in terms of the size of a language model has evolved rapidly (Figure 1). 'Large' in large language models (LLMs) refers to the size of the parameters learned by these language models, which are typically measured in billions, and often trillions, these days. More parameters often allow a model to capture a broader range of linguistic patterns and relationships within the data it was trained on, which can enhance its ability to process and generate texts. Consequently, larger models also generally perform better in

NLP tasks. In 2019, the largest model, to the best of our knowledge, was T5 [10] with up to 11 billion parameters. In comparison, GPT-4 is speculated to have over a trillion parameters. Among open-source alternatives, Meta's LLaMa2 has 70 billion parameters [11]. The utility of LLMs has led to their adoption in biomedical research and application, and some LLMs have been customized or tuned specifically for medical text. In the following sections, we provide an overview of the development and use of LLMs, specifically generative ones, within biomedical NLP. The rest of this chapter is organized as follows: in section 2, we outline the training of biomedical LLMs, currently available biomedical LLMs, and their uses; in sections 3 to 5, we highlight the use of LLMs for distinct types of medical texts—literature, electronic health records (EHRs), and social media, respectively; in section 6, we discuss topics such as bias, reproducibility, and privacy; and we conclude the chapter in section 7 by outlining key limitations and future directions for medical LLMs.

2. Adapting LLMs for the Medical Domain

2.1. Training Medical LLMs

Although generic LLMs like ChatGPT[1] and GPT-4 [12] have demonstrated their robustness in diverse contexts, their performance on medical texts is often suboptimal [13,14], perhaps due to insufficient medical domain-specific training data [15,16]. This gap has catalyzed the emergence of medical-specific LLMs that leverage medical big data to optimize domain-specific performance. Three strategies are typically employed for creating medical domain-specific LLMs [17,18]:

- Pretraining from scratch [19]: This approach requires pretraining a new medical LLM from the ground up using a vast corpus of medical data, employing the transformer architecture [20]. Tailor-made for medical applications, these models can develop a deep and nuanced understanding of medical language and contexts. However, this strategy requires extensive resources in terms of data and computational power;
- Continued pretraining [21]: This strategy involves taking a pre-existing, robust general LLM and further training it on medical-specific data. The key is to infuse the general model with enough medical data so that it can effectively represent/encode and generate medical information. Continued pretraining tasks employ the same strategies as the base models, which can be computationally intensive. Due to the importance and popularity of continued pretraining, particularly for domain adaptation, computationally efficient strategies such as low-rank adaptation (e.g., LoRA) have been proposed [22];
- Instruction tuning [23]: This method fine-tunes general LLMs for the medical domain using instruction tuning data, improving model performance for medical tasks by providing specific instructions. The training data might include task-based instructions, scenario-based queries, or decision-making processes in a medical context. The effectiveness of instruction tuning heavily depends on the quality and variety of the instructional data.

2.2. Emergence of Medical LLMs

Encoder-only transformer models like BERT [2] led the way to the current generative models we refer to as LLMs. Early adaptation of transformer models to the medical domain came through the release of models like BioBERT [7], PubMedBERT [38], and ClinicalBERT [39]. BioBERT was pretrained on the original BERT model using 18 billion words from PMC and PubMed, PubMedBERT was pretrained from a randomly initialized BERT model and using the entirety of PubMed data, and ClinicalBERT [39] was pretrained on a multi-center EHR dataset of diabetes patients. Explorations of the scaling effects for BERT and GPT-2-style models with increasingly large numbers of model weights, up to 8.3 billion, demonstrated that performance increased for both model types at extremely large sizes [40]. This work was followed by pretraining

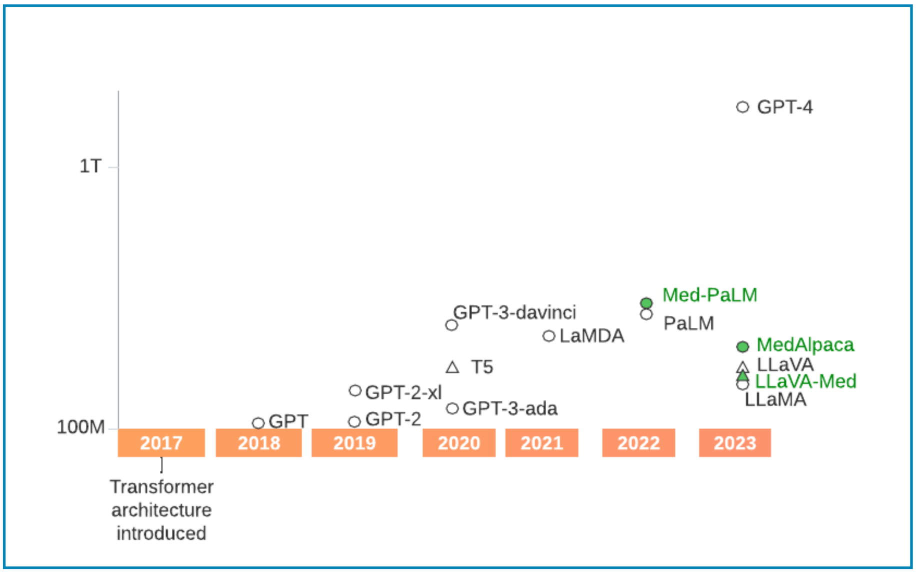

Figure 1. Emergence of larger large language models (LLMs) over time. The vertical axis (parameter size) is represented in the log scale. For models with multiple size variants, only the smallest and largest are shown, for visual clarity. LLMs in green depict those pretrained on biomedical datasets. Triangles are used to depict generative models with an encoder-decoder architecture. All other generative models are decoder-only.

[1] https://openai.com/blog/chatgpt [accessed: 29 Jan, 2024]

using PubMed data from newly initialized models and models pretrained on general corpora [41]. The resulting BioMegatron 345M parameter model showed stronger performance than PubMedBERT on multiple standardized NLP tasks. These works paved the way for new-age generative LLMs with billions of parameters.

We summarize representative LLMs in Table 1. Among them, MedPaLM [24], MedPaLM2 [42], ChatDoctor [25], MedAlpaca [26], Clinical Camel [28], AlpaCare [34], LLaVA-Med [31], MMedLM 2 [37] and Med-Flamingo [33] adopted the instruction tuning strategy. MedPaLM and MedPaLM2, which utilize the PaLM architecture with 540 billion parameters, and PaLM2 as the backbone, have been shown to be particularly effective in medical question-answering (QA) tasks. AMIE [34] also uses PaLM2 as the backbone model, and is specifically fine-tuned for clinical applications, based on medical QA and clinical texts. However, its large parameter size and closed-source nature make it challenging for widespread deployment, particularly in resource-constrained settings. Models such as ChatDoctor and MedAlpaca, based on the open-sourced LLM LLaMA [11, 43], offer a more practical balance. With a focus on QA and conversations, they provide versatility while being less computationally intensive, making them more accessible for diverse medical applications.

GatorTron [44] and GatorTronGPT [29] represent the few medical LLMs that have been trained from scratch. GatorTronGPT is based on the GPT-3 [45] framework, and it was trained using 200 million clinical notes, and 124 NVIDIA DGX nodes.

PMC-LLaMA [15], Clinical-LLaMA [28], Meditron [16] and Me LLaMA [36] represent models using the continued pretraining, or domain-adaptive pretraining approach. These models utilize foundation LLMs that are further enhanced with specific medical knowledge. All these models use LLaMA2 as their backbone, given its strong performance in general domain tasks and open-source nature. PMC-LLaMA and Me LLaMA further combine this approach with instruction tuning, demonstrating an effective, hybrid strategy to adapt the LLaMA2 model to the medical domain.

2.3. Characteristics of Medical LLMs

In the dynamic landscape of medical LLMs, a nuanced understanding of their varied characteristics is crucial for effective application. The distinctions between open and closed-source models, parameter size, and training strategies offer insights into their adaptability and utility in different medical settings.

Open vs. Closed Source. The contrast between open-source models such as ChatDoctor and MedAlpaca and closed-source models like MedPaLM presents a fundamental choice in the medical community. Open-source models foster collaboration, enabling extensive enhancement and personalization. Conversely, closed-source models may offer better professional support, user interfaces, and performance, but confine their development within proprietary boundaries, thus restricting their availability and flexibility.

Parameter Size and Computational Complexity. Parameter size is a pivotal factor in the selection of medical LLMs. Large models like MedPaLM (540 billion parameters) offer in-depth understanding and complex reasoning capabilities. However, their deployment demands substantial resources, making them less feasible for constrained environments. Conversely, smaller models in the 7B-13B range, such as MedAlpaca and AlpaCare, strike a balance between efficiency and effectiveness, facilitating easier deployment.

Domains, Modality, Language and Usage. In the realm of medical LLMs, the majority of studies focus on published biomedical literature text, often overlooking clinical data and tasks. For example, models like PMC-LLaMA and AlpaCare excel in the biomedical literature domain with robust instruction-following capabilities. Meditron, though high-performing, lacks the instruction-following ability, limiting its scope in instruction-driven tasks. Relatively fewer models, such as Clinical-LLaMA and GatorTronGPT, focus on the clinical domain. GatorTronGPT, trained on 82 billion clinical texts and available in 5B and 20B parameter model sizes, demonstrates potential, yet its generalization abilities are constrained by its size. Meanwhile, Clinical-LLaMA, primarily trained on MIMIC-IV's limited clinical texts, concentrates on classification tasks, not fully exploiting the capabilities of LLMs in a variety of clinical settings. Most medical LLMs focus on textual data, but emerging models like Med-Flamingo introduce multimodal capabilities (e.g., via fine-tuning on image caption data), incorporating image analysis which is invaluable in fields like radiology or pathology. The language focus is predominantly English, with exceptions like HuatuoGPT, which is tailored for Chinese, and MMedLM 2, which supports six major languages: Japanese, Spanish, French, Russian, English, and Chinese, indicating a growing trend towards accommodating linguistic diversity in medical LLMs. The choice between these models should be based on the specific needs of biomedical versus clinical applications, data modality, and the complexity of tasks involved.

2.4. Evaluation of LLMs on Biomedical Tasks

The Biomedical Language Understanding Evaluation (BLUE) framework was proposed to facilitate a biomedical counterpart to the popular General Language Understanding Evaluation (GLUE) framework [46]. Comprising five tasks on ten biomedical corpora, resources from BLUE have been partially adopted by some models as part of their evaluation [47,48]. Since the tasks on which medical LLMs are trained are diverse, the evaluation metrics vary. Due to this, there is currently no objective measure to help select an LLM for a specific medical task beyond choosing models that have been specifically pre-trained on domain-specific text. Human evaluation of model outputs has been used in addition to general language model evaluation metrics like BLEU and ROUGE [49,50] for medical text summarization. Similarly, concept extraction from EHRs has been studied through qualitative evaluation [51]. Akin to the BLUE framework, the MultiMedQA framework comprising nine tasks for quantitative evaluation in addition to additional

Table 1. A detailed comparison of medical LLMs available at the time of writing. This comparative analysis spans multiple dimensions, providing a holistic view of each model's characteristics and capabilities. Key aspects of comparison include: backbone model, model size, training strategy, domain-specific data used and data size, accessibility, language, and release date. IFT: instruction fine-tuning; CPT: continued pretraining; SPT: pretraining from scratch; RLMF: reinforcement learning with mixed feedback; QA: question answering. For MedPaLM2, there is no information about the data size and model size of its backbone model PaLM2 (the same backbone model for AMIE). Models without explicit license terms that can be accessed are denoted by a checkmark while models without access are denoted by 'X'. Consumer-Mediated Health Information Exchange: Authorize Direct Transmission.

Model	Backbone	Size	Strategy	Data	Data Size	Modality	License Access	Language	Release date
MedPaLM [24]	PaLM	540B	IFT	QA	-	Text	X	English	12/26/22
ChatDoctor [25]	LLaMA	7B	IFT	Conversation	100K	Text	Apache 2.0	English	03/24/23
MedAlpaca [26]	LLaMA	7B, 13B	IFT	QA	160K	Text	GNU GPL v3.0	English	04/14/23
PMC-LLaMA [15]	LLaMA, LLaMA2	7B, 13B	CPT+IFT	Literature, book, conversation, QA, knowledge graph	79B, 514K	Text	Apache 2.0	English	04/25/23
MedPaLM2 [27]	PaLM2	-	IFT	QA	-	Text	X	English	05/16/23
Clinical Camel [28]	LLaMA2	13B, 70B	IFT	Conversation, QA	104K	Text	GNU Affero GPL v3.0	English	05/19/23
GatorTronGPT [29]	GPT-3 architecture	5B, 20B	SPT	General and clinical text	82B	Text	Apache 2.0	English	05/24/23
HuatuoGPT [30]	Baichuan, Ziya	7B, 13B	IF-T+RLMF	Conversation, QA	-	Text	Apache 2.0	Chinese	05/25/23
LLaVA-Med [31]	LLaVA	13B	IFT	biomedical image-text	630K	Text, image	Microsoft ResearchLicense	English	06/01/23
Clinical-LLaMA [32]	LLaMA2	7B	CPT	MIMIC-IV	-	Text	X	English	07/12/23
Med-Flamingo [33]	Flamingo	9B	IFT	General and biomedical image-text	>0.01B	Text, image	√	English	07/27/23
AlpaCare [34]	LLaMA2	7B, 13B	IFT	Biomedical conversation	52K	Text	Apache 2.0	English	10/23/23
Meditron [16]	LLaMA2	7B, 70B	CPT	General and biomedical text	48B	Text	Apache 2.0	English	11/27/23
AMIE [35]	PaLM2	-	IFT	QA, clinical text	110K	Text	X	English	01/11/24
Me LLaMA [36]	LLaMA2	13B, 70B	CPT+IFT	QA, clinical text	129B, 214K	Text	PhysioNet Credentialed HealthData License 1.5.0	English	02/20/24
MMedLM2 [37]	InternLM, BLOOM	7B	IFT	QA, clinical text	25.5B	Text	cc-by-4.0	Multilingual	02/26/24

tasks for qualitative evaluation has been proposed [52]. The MedEval testbed focuses on covering a wider variety of human body parts, as compared to existing frameworks that prioritize task coverage [53].

Beyond the English language, MedBench [54] and MMedBench [37] have been introduced as evaluation frameworks for Chinese and multilingual medical LLMs, respectively. BioMistral[2], a collection of open-source pretrained biomedical LLMs, also features a multilingual evaluation framework in seven languages, although the drawbacks of auto-translation remain as they translate an English benchmark into seven languages. The framework CRAFT-MD[3] focuses exclusively on the task of conversational reasoning for evaluating medical LLMs. Since evaluation of LLMs is an evolving area of research, just like LLMs themselves, novel evaluation strategies (such as hallucination evaluation [55]) are actively being proposed, and we anticipate a trend of more standardized and converging benchmarking of medical LLMs in the future.

3. LLMs for Medical Literature

Most training and application of medical LLMs have focused on text from the medical literature. In this section, we briefly review key NLP tasks, highlight some of the most important models that were trained on the biomedical literature, discuss their applications to specific tasks, and address challenges associated with their use.

3.1. LLM Applications in Biomedical Literature

Several NLP tasks have relevance when applying LLMs to the biomedical literature. Named Entity Recognition (NER) is the identification of words that belong to a class of interest (drugs, symptoms, etc.). Relation extraction is identifying relationships between named entities from the literature (e.g., drug A treats condition Z). Literature-based discovery is the prediction of new relationships between entities using existing entities and relationships extracted from a corpus. Other tasks include question answering (QA) where the model responds to a query given some biomedical text as input, information retrieval to identify and extract relevant information from text, and summarization to provide a concise description of larger biomedical text.

3.1.1. Information Extraction

LLMs have shown moderate to good, though not state-of-the-art, performance on NER. For example, a GPT model augmented with a biomedical knowledge graph was able to obtain respectable F1 scores [56]. The same study also showed that LLMs were able to achieve impressive scores for zero-shot NER. There is some potential for LLMs to be used for drug repurposing, a literature-based discovery task, as recent research has demonstrated that even the generic GPT-4 model can perform well on this task [57] by fine-tuning the model with prompts that are augmented with knowledge from a biomedical knowledge graph. Recently, there have been explorations into using LLMs, such as MedLLaMA, for relation extraction and knowledge graph construction [58]. While LLMs can be used for information extraction, a recent survey found that PubMedBERT often outperformed GPT-3.5 and GPT-4 on information retrieval-focused tasks [59]. These studies demonstrated that there is room for improvement on these tasks for LLMs.

3.1.2. Question Answering

LLMs have demonstrated exceptional performance on a wide variety of QA-focused tasks. ChatDoctor, PMC-LLaMA, MedAlpaca [26] have all demonstrated strong performance on a range of biomedical QA datasets. This suggests that these models, tuned on the literature, may be useful for information retrieval and for reducing the time physicians need to spend looking through PubMed to find answers to their questions. MedAlpaca, in particular, has achieved state-of-the-art zero-shot performance in answering questions from the three parts of the United States Medical Licensing Exam series [26].

3.1.3. Information Retrieval

Combining LLMs with external databases for information retrieval and response generation, retrieval augmented generation, has shown promise in improving the specificity of answers provided to users [60]. As discussed later in this survey paper, hallucinations are a problem with any LLM but can be particularly hazardous in a medical context where a hallucination could lead to erroneous advice generation. However, this can be mitigated by the use of knowledge graphs to enhance reasoning and ground LLMs to decrease the likelihood of hallucinations [57].

3.1.4. Summarization

Recent work has shown that LLMs can effectively be fine-tuned on small amounts of domain-specific data to produce high-quality summarizations of biomedical research abstracts [61]. A sample of 175 abstracts from the specialized COVID-19 dataset, CORD-19 [62], was used with GPT-3.5 and Davinci (the initial model of the GPT-3 series) to produce a dataset of synthetic prompts that were used to successfully fine-tune other LLMs to produce summarizations of abstracts.

4. LLMs for Electronic Health Records

Large-scale data mining from free text notes in EHRs and extracting meaningful information has been a challenge for NLP research over the years, and LLMs present substantial opportunities to move the state-of-the-art forward. Consequently, LLMs have found widespread use in the context of EHRs. Here, we outline a subset of key applications of LLMs for EHR data.

[2] https://arxiv.org/abs/2402.10373

[3] https://www.medrxiv.org/content/10.1101/2023.09.12.23295399v2

4.1. Clinical Decision Support Systems

Clinical Decision Support Systems (CDSSs) enhance medical decision-making by providing targeted clinical knowledge, patient information, and other health information [63,64]. Recently, LLMs have shown great potential as core CDSS components. LLMs trained or fine-tuned on texts from EHRs can capture high-quality representations of input texts that can be used for tasks such as diagnosis prediction [65], clinical trial outcome measurement [66], and in-hospital mortality prediction [67], to name a few. In-context learning capabilities [6,68] make generative LLMs inherently interactive, allowing for direct conversations and QA sessions between physicians and LLMs. Such interactive artificial intelligence (AI) systems built on NLP technology was largely infeasible prior to the recent advances in generative LLMs. Chatbots have become a common component in new-age CDSS designs with generative LLM integration [69,70]. A number of studies have already demonstrated the effectiveness of using such chatbots for clinical decision support, particularly illustrating their potential to improve productivity in healthcare settings and facilitate better health outcomes [29,71,72]. While research in this space is still early, the overwhelmingly promising results suggest that chatbots based on generative LLMs and trained on EHR texts will see significant adoption across healthcare systems. In terms of specific applications, text-to-text LLMs have found use in chest X-ray report summarization [73], summarization and classification of medical dialogues [74], and diagnostic reasoning [75], to name a few. Due to the many possible applications of LLMs on EHR texts, as mentioned earlier in this article, a number of models have been proposed that have been trained or fine-tuned on EHR-specific data—the most notable of them perhaps being GatorTron [44].

4.2. Revenue Cycle Optimization

Revenue Cycle Management (RCM) is a critical process for healthcare organizations that typically consists of pre-encounter, intra-encounter, and post-encounter steps [76]. Revenue Cycle Optimization (RCO) analyzes each of these steps to identify areas that can increase revenue, reduce expenses, and improve cash flow. Therefore, RCO helps streamline RCM, resulting in efficient billing processes, reduced claim denials, and timely reimbursements, which can significantly improve the overall quality of care, ultimately driving better health outcomes [77]. Traditionally, RCO is a manual, highly complex endeavor with plenty of opportunities to augment human decision-making with AI.

Generative LLM-based chatbots can help with pre-encounter and inter-encounter tasks, including appointment scheduling, patient registration, and prior authorization. Given the politeness and empathy that can be infused into these chatbots in simulated healthcare settings [78], their use in patient interactions holds great promise. According to a 2019 report [79], payment errors amount to hundreds of billions of dollars in unnecessary spending annually. To this end, NYUTron has shown that EHR-trained generative LLMs can predict insurance denials, reducing billing complexity and payment errors [80].

4.3. Translational Research

Translational research is defined as "the process of applying ideas, insights and discoveries generated through basic scientific inquiry to the treatment or prevention of human disease" [81]. The application of AI-driven methods to translational research presents a clear opportunity for multidisciplinary collaborations, with teams comprising diverse domain experts, such as computer scientists, mathematicians, and clinicians. Indeed, many successful translational research efforts have had authors with wide-ranging research backgrounds [82-84]. This trend is likely to continue in the case of LLMs trained on EHRs, with researchers discovering novel applications to medicine [85]. Clinical trials, the gold standard for evaluating new treatments, are critical in translational research. Feasibility studies before conducting clinical trials leverage LLMs to analyze patient records, aiding in the assessment of a trial's practicality by quickly determining if sufficient suitable participants are available. This accelerates the early stages of research. LLMs can also analyze EHRs to extract and normalize medical concepts to identify potential trial participants. Systems like EliIE [86] could be enhanced by LLMs, facilitating patient cohort definition and even enabling in silico trials. Moreover, LLMs can play a crucial role in patient recruitment, swiftly analyzing clinical documents in EHRs to match patients with appropriate trials, thereby streamlining recruitment and ensuring that trial opportunities are maximized. Lastly, LLMs can aid in analyzing clinical trial criteria and matching patients with the most suitable trials, offering patients access to the latest treatment options, for example, TrialGPT [87].

5. LLMs for Health-related Social Media Data

Social media chatter contains an abundance of health-related information as subscribers often discuss such topics with their peers. A small survey, for example, showed that over 80% of cancer patients conducted disease-related communication with others over social media [88]. Often, health-related information available from social media are not available from any other sources [89]. Knowledge relevant to health can be acquired from social media in close to real-time [90], at scale, and often from hard-to-reach populations [91]. NLP of health-related social media texts has proven difficult (relative to, for example, text from medical literature) over the years for various data-level characteristics, such as misspellings and colloquial and evolving language [92]. Despite the difficulties associated with NLP of social media data, recent progress has demonstrated their utility for complex tasks, such as targeted public health surveillance [93] and behavior analysis [94].

Literature on the application of generative LLMs for social media based health-related tasks is sparse, with only a handful of papers published to date. This is unsurprising since applications of new NLP technology within the health space typically target domain-specific text from sources other than social media such as medical literature. Once these methods are mature, they are applied and often customized for social media texts, which represent an array of additional NLP challenges. We outline some important recent research contributions, noting that most language models in application today are still transformer-based encoder models, which outperform new-age generative models in benchmarking experiments [95].

5.1. LLMs for Mental Health

Social media sources encapsulate a trove of knowledge regarding mental health, and, consequently, a substantial chunk of health-related NLP research involving social media data has focused on mental health-related topics [96]. With the emergence of new-age LLMs, the trend has been no different. A very recent contribution is the MentaLLaMA model [97]. The paper introducing the model also presents the interpretable mental health instruction dataset built from social media data. The MentalLLaMA model is shown to perform close to supervised state-of-the-art discriminative models while providing reasonable explanations. Despite less than optimal performance, MentaLLaMA represents an exciting branching of open-source LLMs—the first to specifically address health-related social media text. A similar theme of using social media data for providing conversational text assistance pertaining to mental health issues is demonstrated by Psy-LLM [98]. They use data from the Chinese social media platforms Tianya, Yixinli, and Zhihu. Inevitably, many more models customized for targeted health-related topics over social media data will follow suit.

Xu et al. [99] present a comparative evaluation of zero-shot, few-shot (examples inserted into the prompt), and instruction fine-tuned LLMs on mental health datasets obtained from the social media platform Reddit. Their fine-tuned models Mental-Alpaca, based on Alpaca [100], and Mental-FLANT5, based on the T5 framework [10], outperform GPT-3.5 and GPT-4. These results demonstrate the promise of subdomain-specific LLMs in digital health.

5.2. LLMs for Health Surveillance

Identifying and monitoring disease outbreaks in real-time with social media has been useful for health institutions, including the World Health Organization [101]. Although the use of LLMs in infoveillance is still nascent, the ability of LLMs to extract symptoms from social media posts has already been explored [102] in the context of COVID-19. The study showed that with prompt engineering, GPT-3.5-Turbo and GPT-4 can achieve high precision and recall. The authors also reported using ChatGPT as an aid for prompt engineering to arrive at the final prompt used to compare different LLMs, including Google Bard. Such use of LLMs to aid in using other LLMs is likely to rise in the future.

5.3. LLMs for Combating Health Misinformation

The rapid spread of misinformation through social media is an active topic of concern among the broader healthcare community [101]. With LLMs generating text that is indistinguishable from those written by humans, there is notable risk of misuse of LLMs to generate misinformation at scale. Chen et al. [103] point out the dichotomy of LLM usage in both promoting as well as combating health-related misinformation on social media platforms. They highlight the use cases of detecting health-related misinformation by augmenting LLMs with external knowledge bases. With social media platforms integrating LLMs more closely into their ecosystem, misinformation intervention techniques such as LLM-generated factual counter-misinformation responses are on the horizon. Xiao et al. [104] take a step in this direction by creating Jennifer, a chatbot to provide expert-sourced answers to commonly asked questions about COVID-19. They deployed their chatbot on Facebook, in addition to other websites, and found that Jennifer was able to help users find more accurate answers to COVID-19-related queries. They proposed using LLMs to complement their expert-sourcing framework for faster deployment in future versions of their chatbot.

5.4. LLMs for Health Opinion Mining

The massive amount of opinion generated on social media platforms makes them lucrative for gauging public sentiments toward health policies as well as early detection of rapidly changing health threats like COVID-19. Tran et al. [105] utilized LLMs for COVID-19-related public opinion mining using Twitter (X) data from Japan. Although the LLMs exhibited high performance variability in zero-shot settings with different prompt designs, the study showed a useful application of LLMs in digital health.

6. LLMs: Challenges, Limitations, and Risks

6.1. Privacy and Data Security

Closed-source LLMs such as ChatGPT have seen substantial rise in popularity owing to their ease of use. These LLMs invoke API[4] calls to the model hosted off-site with possible personally identifiable information (PII) being transmitted. Some healthcare institutions have established internal guidelines to prevent such off-site transmission of data, while some have reached agreements with service providers to securely transfer PII. Much like LLMs themselves, privacy and data security standards associated with LLM use in healthcare settings are evolving. Mesko and Topol [106] highlight the need for informed patient consent for the use of LLMs in their personal healthcare management.

[4] Application Programming Interface; for communication between a server and the client/user.

6.2. Bias and Fairness

Several studies have found that LLMs do not generalize well across distinct demographics. GPT-4 has been shown to exhibit bias when diagnosing people of different genders, races, and ethnicities [107]. Regulation of LLMs, particularly in healthcare, has been suggested [106] as a means to protect vulnerable populations from harm. The sparsity of multilingual medical LLMs is another factor contributing to geographical bias in the adoption of medical LLMs. The MMedC medical corpus covering five languages, CHIMED-GPT for Chinese, and the bilingual mixture-of-experts LLM BiMedX for Arabic medical tasks are noteworthy attempts to extend the application of LLMs in more languages. Perhaps the most impactful stride toward alleviating geographical disparities is demonstrated by Apollo [108]. Covering a population of 6 billion through six languages, Apollo is further distinguished by being lightweight.

6.3. Interpretability, Explainability, Transparency, and Equitable AI

A major issue hindering the adoption of LLM-based technologies in healthcare is the lack of transparency. LLMs may exhibit opacity across different dimensions: closed-source models, lack of information on training data and procedures, and limited interpretability and explainability. This lack of transparency adds to the issue of accountability, which is paramount in medicine. Although interventions such as the foundation model transparency index [109] have been proposed, the success of such approaches largely relies on their adoption.

The compute-intensive nature of large models poses challenges to equitable use of AI. For large models, it is necessary to utilize clusters of GPUs for time-efficient fine-tuning and inference. For example, at full precision, LLaMa-v2 70B requires at least 8 A100 NVDIA GPUs for tuning. Not all researchers and developers have access to the resources necessary to tune or deploy these models, which inevitably leads to inequity in research and application of LLMs in medicine. Some techniques in “TinyML” have been proposed to reduce the footprint of these large models and include techniques such as training subsets of weights (LoRA), quantization, and precision reduction [110].

6.4. Evaluation of LLM Performance, Reproducibilty, and Deployability

In terms of NLP tasks, multiple studies have shown that for tasks such as supervised classification and information extraction, transformer-based models still outperform LLMs when large labeled training data is available. LLMs, however, typically perform better for these tasks in low-/zero-shot settings [56,111]. For generative tasks such as summarization and QA, LLMs have demonstrated the greatest improvements compared to earlier methods [75, 112]. Xu et al. [99] stress, in the context of mental health, that despite the impressive performance of LLMs, their deployability for digital health applications remains in question. For deployment in healthcare, LLMs must be evaluated on aspects beyond accuracy or other intrinsic evaluation metrics. Targeted pre- and post-deployment evaluations of fairness, bias, and interpretability are essential, along with robust strategies for effective identification and mitigation of ethical concerns [107].

Reproducibility can often be challenging to ensure since LLM outputs may be non-deterministic. This may particularly happen with API-based LLMs since the closed-source back-end models may change over time, rendering past outputs irreproducible. Additionally, open-source models may lose their initial capabilities when further tuned on new data, a problem known as catastrophic forgetting [21,113,114].

Finally, LLMs used for generative tasks pose a risk for hallucinations that can be at best annoying or at worst dangerous in the case of biomedical text generation for QA or summarization. Hallucinations can be particularly hard to predict and may remain undiscovered if models are not evaluated rigorously. The ease of use of API-based models like GPT-4 particularly make non-expert researchers (i.e., those without substantial experience in NLP research) susceptible to interpreting coherent responses generated by LLMs to be accurate information. Extensive intrinsic and extrinsic evaluations, led by researchers with substantial NLP expertise, are necessary to mitigate the dangers of hallucination.

7. Concluding Remarks

Notwithstanding the challenges associated with the use of LLMs in biomedical NLP tasks, there remains vast potential for integrating LLMs in research, education, and clinical care [115]. Retrieval augmented generation has been gaining popularity as a method to reduce hallucinations [116].

Prompt modification, integration of external data sources, and leveraging knowledge graphs for reasoning augmentation have also been found to mitigate hallucinations [117]. Evaluating general purpose LLMs for trustworthiness across several dimensions such as reliability, safety, and resistance to misuse has been proposed [118]. Designing trustworthiness criteria specific to healthcare systems can benefit the advancement of medical LLMs. As more effective approaches for mitigating the limitations associated with LLMs are developed, their adoption in biomedical NLP tasks is expected to increase further. We expect to see a trend of more equitable dissemination of LLM infrastructure with smaller, less resource-intensive models in the future. Biomedical subdomain-specific models, similar to MentalLlaMA for the subdomain of mental health, are also envisaged.

LLMs have brought about a new paradigm in biomedical NLP research and application, enabling us to make substantial progress on problems that appeared unsolvable even in the recent past (e.g., QA). Thus, the proliferation of LLM use in digital health is largely desirable, and the future is promising. However, it is critical to establish and implement necessary guardrails for responsible usage to mitigate the possibility of unintended harm.

References

1. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Represen-tations of Words and Phrases and their Compositionality. In: Burges CJ, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc.; 2013. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
2. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR. 2018;abs/1810.04805. http://arxiv.org/abs/1810.04805
3. Mahajan D, Liang JJ, Tsou CH, Ozlem Uzuner. Overview of the 2022 n2c2 shared task on contex-tualized medication event extraction in clinical notes. Journal of Biomedical Informatics. 2023;144:104432. https://doi.org/10.1016/j.jbi.2023.104432
4. Klein AZ, Banda JM, Guo Y, Schmidt AL, Xu D, Amaro IF, et al. Overview of the 8th social media mining for health applications (SMM4H) shared tasks at the AMIA 2023 annual symposium. Journal of the American Medical Informatics Association. 2024 01:ocae010. https://doi.org/10.1093/jamia/ocae010
5. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc.; 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf
6. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in Neural Information Processing Systems. vol. 33. Curran Associates, Inc.; 2020. p. 1877-901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
7. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language rep-resentation model for biomedical text mining. Bioinformatics. 2020 Feb;36(4):1234-40. http://doi.org/10.1093/bioinformatics/btz682
8. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Rumshisky A, Roberts K, Bethard S, Naumann T, editors. Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72-8. https://aclanthology.org/W19-1909
9. Min B, Ross H, Sulem E, Veyseh APB, Nguyen TH, Sainz O, et al. Recent advances in natural lan-guage processing via large pre-trained language models: A survey. ACM Computing Surveys. 2023;56(2):1-40. https://doi.org/10.1145/3605943
10. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. CoRR. 2019;abs/1910.10683. http://arxiv.org/abs/1910.10683
11. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023. https://arxiv.org/abs/2307.09288
12. OpenAI. GPT-4 Technical Report; 2023. https://arxiv.org/abs/2303.08774
13. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in Artificial Intelligence. 2023;6:1169595. https://doi.org/10.3389/frai.2023.1169595
14. Sallam M, Salim N, Barakat M, Al-Tammemi A. ChatGPT applications in medical, dental, pharmacy, and public health education: A descriptive study highlighting the advantages and limitations. Narra J. 2023;3(1):e103-3. https://doi.org/10.52225/narra.v3i1.103
15. Wu C, Zhang X, Zhang Y, Wang Y, Xie W. PMC-LLaMA: Towards Building Open-source Language Models for Medicine. arXiv preprint arXiv:230414454. 2023. https://arxiv.org/abs/2304.14454
16. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. arXiv preprint arXiv:231116079. 2023. https://arxiv.org/abs/2311.16079
17. Wang B, Xie Q, Pei J, Chen Z, Tiwari P, Li Z, et al. Pre-trained language models in biomedical do-main: A systematic survey. ACM Computing Surveys. 2023;56(3):1-52. https://doi.org/10.1145/3611651
18. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. arXiv pre-print arXiv:230318223. 2023. https://arxiv.org/abs/2303.18223
19. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pre-training for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021;3(1):1-23. https://doi.org/10.1145/3458754
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30. https://papers.nips.cc/paper/7181-attention-is-all-you-need
21. Ke Z, Shao Y, Lin H, Konishi T, Kim G, Liu B. Continual Pre-training of Language Models. In: The Eleventh International Conference on Learning Representations; 2022. https://doi.org/10.18653/v1/2022.emnlp-main.695
22. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al. LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations; 2022. https://openreview.net/forum?id=nZeVKeeFYf9
23. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023 09;330(9):866-9. https://doi.org/10.1001/jama.2023.14217
24. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. arXiv preprint arXiv:221213138. 2022. https://arxiv.org/abs/2212.13138
25. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. Cureus. 2023;15(6). https://doi.org/10.7759/cureus.40895
26. Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca– An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv preprint arXiv:230408247. 2023. https://arxiv.org/abs/2304.08247
27. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. arXiv preprint arXiv:230714334. 2023. https://arxiv.org/abs/2307.14334
28. Toma A, Lawler PR, Ba J, Krishnan RG, Rubin BB, Wang B. Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. arXiv preprint arXiv:230512031. 2023. https://arxiv.org/abs/2305.12031
29. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A Study of Generative Large Language Model for Medical Research and Healthcare. npj Digital Medicine. 2023 Nov;6(1):210. Available from: https://doi.org/10.1038/s41746-023-00958-w
30. Zhang H, Chen J, Jiang F, Yu F, Chen Z, Chen G, et al. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In: Findings of the Association for Computational Linguistics: EMNLP 2023; 2023. p. 10859-85. https://doi.org/10.18653/v1/2023.findings-emnlp.725
31. Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:230600890. 2023. https://arxiv.org/abs/2306.00890
32. Gema A, Daines L, Minervini P, Alex B. Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain. arXiv preprint arXiv:230703042. 2023. https://arxiv.org/abs/2307.03042
33. Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, et al. Med-Flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). PMLR; 2023. p. 353-67. https://proceedings.mlr.press/v225/moor23a.html
34. Zhang X, Tian C, Yang X, Chen L, Li Z, Petzold LR. AlpaCare: Instruction-tuned Large Language Models for Medical Application. arXiv preprint arXiv:231014558. 2023. https://arxiv.org/abs/2310.14558
35. Tu T, Palepu A, Schaekermann M, Saab K, Freyberg J, Tanno R, et al. Towards Conversational Diagnostic AI. arXiv preprint arXiv:240105654. 2024. https://arxiv.org/abs/2401.05654
36. Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, et al. Me LLaMA: Foundation Large Language Models for Medical Applications. arXiv preprint arXiv:240212749. 2024. https://arxiv.org/abs/2402.12749

37. Qiu P, Wu C, Zhang X, Lin W, Wang H, Zhang Y, et al. Towards Building Multilingual Language Model for Medicine. arXiv preprint arXiv:240213963. 2024. https://arxiv.org/abs/2402.13963
38. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pre-training for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare. 2022 Jan;3(1):1-23. https://doi.org/10.1145/3458754
39. Wang G, Liu X, Ying Z, Yang G, Chen Z, Liu Z, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. Nature Medicine. 2023;29(10):2633-42. https://doi.org/10.1038/s41591-023-02552-9
40. Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. arXiv; 2020. ArXiv:1909.08053. http://arxiv.org/abs/1909.08053
41. Shin HC, Zhang Y, Bakhturina E, Puri R, Patwary M, Shoeybi M, et al. BioMegatron: Larger Bio-medical Domain Language Model. arXiv; 2020. ArXiv:2010.06060. http://arxiv.org/abs/2010.06060
42. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. https://doi.org/10.1038/s41586-023-06291-2
43. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:230213971. 2023. https://arxiv.org/abs/2302.13971
44. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, et al. A large language model for electronic health records. npj Digital Medicine. 2022;5(1). https://doi.org/10.1038/s41746-022-00742-2
45. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
46. Peng Y, Yan S, Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: Proceedings of the 2019 Workshop on Bi-omedical Natural Language Processing (BioNLP 2019); 2019. https://doi.org/10.18653/v1/W19-5006
47. Zhang K, Yu J, Yan Z, Liu Y, Adhikarla E, Fu S, et al. BiomedGPT: A unified and generalist biomedi-cal generative pre-trained transformer for vision, language, and multimodal tasks. arXiv preprint arXiv:230517100. 2023. https://arxiv.org/abs/2305.17100
48. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pretrained transformer for biomedical text generation and mining. Briefings in bioinformatics. 2022;23(6):bbac409. https://doi.org/10.1093/bib/bbac409
49. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. npj Digital Medicine. 2023;6(1):158. https://doi.org/10.1038/s41746-023-00896-7
50. Singh J, Patel T, Singh A. Performance Analysis of Large Language Models for Medical Text Sum-marization. OSF Preprints. https://osf.io/preprints/osf/kn5f2
51. Liu D, Ding C, Bold D, Bouvier M, Lu J, Shickel B, et al. Evaluation of General Large Language Models in Contextually Assessing Semantic Concepts Extracted from Adult Critical Care Electronic Health Record Notes. arXiv preprint arXiv:240113588. 2024. https://arxiv.org/abs/2401.13588
52. Abraham TM, Adams G. Evaluating the Medical Knowledge of Open LLMs - Part 1. MedARC Blog. 2024 January. https://medarc.ai/blog/medarc-llms-eval-part-1
53. He Z, Wang Y, Yan A, Liu Y, Chang E, Gentili A, et al. MedEval: A Multi-Level, Multi-Task, and Multi-Domain Medical Benchmark for Language Model Evaluation. In: Proceedings of the 2023 Con-ference on Empirical Methods in Natural Language Processing; 2023. p. 8725-44. https://doi.org/10.18653/v1/2023.emnlp-main.540
54. Cai Y, Wang L, Wang Y, de Melo G, Zhang Y, Wang Y, et al. Medbench: A large-scale chinese benchmark for evaluating medical large language models. arXiv preprint arXiv:231212806. 2023. https://arxiv.org/abs/2312.12806
55. Zhu Z, Sun Z, Yang Y. HaluEval-Wild: Evaluating Hallucinations of Language Models in the Wild. arXiv preprint arXiv:240304307. 2024. https://arxiv.org/abs/2403.04307
56. Bian J, Zheng J, Zhang Y, Zhu S. Inspire the Large Language Model by External Knowledge on Bi-oMedical Named Entity Recognition. arXiv; 2023. ArXiv:2309.12278. http://arxiv.org/abs/2309.12278
57. Soman K, Rose PW, Morris JH, Akbas RE, Smith B, Peetoom B, et al. Biomedical knowledge graph-enhanced prompt generation for large language models. arXiv; 2023. ArXiv:2311.17330. http://arxiv.org/abs/2311.17330
58. Li M, Chen M, Zhou H, Zhang R. PeTailor: Improving Large Language Model by Tailored Chunk Scorer in Biomedical Triple Extraction. arXiv; 2023. ArXiv:2310.18463. http://arxiv.org/abs/2310.18463
59. Chen Q, Du J, Hu Y, Keloth VK, Peng X, Raja K, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv; 2024. ArXiv:2305.16326. http://arxiv.org/abs/2305.16326
60. Wang C, Ong J, Wang C, Ong H, Cheng R, Ong D. Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation. Annals of Biomedical Engi-neering. 52, 1115–1118 (2024). https://doi.org/10.1007/s10439-023-03327-6
61. Khan YA, Hokia C, Xu J, Ehlert B. covLLM: Large Language Models for COVID-19 Biomedical Lit-erature. arXiv; 2023. ArXiv:2306.04926. http://arxiv.org/abs/2306.04926
62. Wang LL, Lo K, Chandrasekhar Y, Reas R, Yang J, Burdick D, et al. CORD-19: The COVID-19 Open Research Dataset. arXiv; 2020. ArXiv:2004.10706. http://arxiv.org/abs/2004.10706
63. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digital Medicine. 2020 Feb;3(1):17. https://doi.org/10.1038/s41746-020-0221-y
64. Osheroff J, Teich J, Levick D, Saldana L, Velasco F, Sittig D, et al. Improving outcomes with clinical decision support. 2nd ed. HIMSS Book Series. Himss Publishing; 2012. https://doi.org/10.4324/9781498757461
65. van Aken B, Papaioannou JM, Naik M, Eleftheriadis G, Nejdl W, Gers F, et al. This Patient Looks Like That Patient: Prototypical Networks for Interpretable Diagnosis Prediction from Clinical Text. In: He Y, Ji H, Li S, Liu Y, Chang CH, editors. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online only: Association for Computational Linguistics; 2022. p. 172-84. https://aclanthology.org/2022.aacl-main.14
66. Lee RY, Kross EK, Torrence J, Li KS, Sibley J, Cohen T, et al. Assessment of Natural Language Pro-cessing of Electronic Health Records to Measure Goals-of-Care Discussions as a Clinical Trial Outcome. JAMA Network Open. 2023 03;6(3):e231204. https://doi.org/10.1001/jamanetworkopen.2023.1204.
67. Lyu W, Dong X, Wong R, Zheng S, Abell-Hart K, Wang F, et al. A Multimodal Transformer: Fusing Clinical Notes with Structured EHR Data for Interpretable In-Hospital Mortality Prediction. AMIA Annu Symp Proc. 2022;2022:719-28. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10148371
68. Garg S, Tsipras D, Liang P, Valiant G. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. In: Oh AH, Agarwal A, Belgrave D, Cho K, editors. Advances in Neural In-formation Processing Systems; 2022. Available from: https://openreview.net/fo-rum?id=flNZJ2eOet.
69. Benary M, Wang XD, Schmidt M, Soll D, Hilfenhaus G, Nassir M, et al. Leveraging Large Language Models for Decision Support in Personalized Oncology. JAMA Network Open. 2023 Nov;6(11):e2343689. http://dx.doi.org/10.1001/jamanetworkopen.2023.43689
70. Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, et al. Using AI-generated sug-gestions from ChatGPT to optimize clinical decision support. Journal of the American Medical In-formatics Association. 2023 04;30(7):1237-45. https://doi.org/10.1093/jamia/ocad072
71. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. J Med Internet Res. 2023 Aug;25:e48659. https://doi.org/10.2196/48659

72. Fawzi S. A Review of the Role of ChatGPT for Clinical Decision Support Systems. In: 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES); 2023. p. 439-42. https://doi.org/10.1109/NILES59815.2023.10296668.
73. Wang T, Zhao X, Rios A. UTSA-NLP at RadSum23: Multi-modal Retrieval-Based Chest X-Ray Re-port Summarization. In: Demner-fushman D, Ananiadou S, Cohen K, editors. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. Toronto, Canada: Associ-ation for Computational Linguistics; 2023. p. 557-66. Available from: https://aclanthology.org/2023.bionlp-1.58
74. Ozler KB, Bethard S. clulab at MEDIQA-Chat 2023: Summarization and classification of medical dialogues. In: Naumann T, Ben Abacha A, Bethard S, Roberts K, Rumshisky A, editors. Proceedings of the 5th Clinical Natural Language Processing Workshop. Toronto, Canada: Association for Computational Linguistics; 2023. p. 144-9. https://aclanthology.org/2023.clinicalnlp-1.19
75. Sharma B, Gao Y, Miller T, Churpek M, Afshar M, Dligach D. Multi-Task Training with In-Domain Language Models for Diagnostic Reasoning. In: Naumann T, Ben Abacha A, Bethard S, Roberts K, Rumshisky A, editors. Proceedings of the 5th Clinical Natural Language Processing Workshop. Toronto, Canada: Association for Computational Linguistics; 2023. p.78-85. https://aclanthology.org/2023.clinicalnlp-1.10
76. Chin S, Li A, Boulet M, Howse K, Rajaram A. Resident and family physician perspectives on billing: An exploratory study. Perspect Health Inf Manag. 2022 Oct;19(4):1g. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9635049/
77. Bhati D, Deogade MS, Kanyal D. Improving Patient Outcomes Through Effective Hospital Admin-istration: A Comprehensive Review. Cureus. 2023 Oct. http://dx.doi.org/10.7759/cureus.47731
78. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. JAMA Internal Medicine. 2023 06;183(6):589-96. https://doi.org/10.1001/jamainternmed.2023.1838
79. Shrank WH, Rogstad TL, Parekh N. Waste in the US Health Care System: Estimated Costs and Po-tential for Savings. JAMA. 2019 10;322(15):1501-9. https://doi.org/10.1001/jama.2019.13978
80. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023 Jun;619(7969):357-62. http://dx.doi.org/10.1038/s41586-023-06160-y
81. Lost in clinical translation. Nature Medicine. 2004 Sep;10(9):879-9. https://doi.org/10.1038/nm0904-879
82. Yala A, Mikhael PG, Lehman C, Lin G, Strand F, Wan YL, et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. Nature Medicine. 2022 Jan;28(1):136-43. https://doi.org/10.1038/s41591-021-01599-w
83. Rehman RZU, Del Din S, Guan Y, Yarnall AJ, Shi JQ, Rochester L. Selecting Clinically Relevant Gait Characteristics for Classification of Early Parkinson's Disease: A Comprehensive Machine Learning Approach. Scientific Reports. 2019 Nov;9(1):17269. https: //doi.org/10.1038/s41598-019-53656-7
84. Guazzo A, Longato E, Fadini GP, Morieri ML, Sparacino G, Di Camillo B. Deep-learning-based nat-ural-language-processing models to identify cardiovascular disease hospitalisations of patients with diabetes from routine visits' text. Scientific Reports. 2023 Nov;13(1):19132. https://doi.org/10.1038/s41598-023-45115-1
85. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language mod-els in medicine. Nature Medicine. 2023 Aug;29(8):1930-40. https://doi.org/10.1038/s41591-023-02448-8
86. Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, et al. EliIE: An open-source infor-mation extraction system for clinical trial eligibility criteria. Journal of the American Medical In-formatics Association. 2017;24(6):1062-71. https://doi.org/10.1093/jamia/ocx019
87. Jin Q, Wang Z, Floudas CS, Sun J, Lu Z. Matching patients to clinical trials with large language models. ArXiv. 2023. https://arxiv.org/abs/2307.15051
88. Braun LA, Zomorodbakhsch B, Keinki C, Huebner J. Information needs, communication and usage of social media by cancer patients and their relatives. Journal of cancer research and clinical on-cology. 2019;145:1865-75. https://doi.org/10.1007/s00432-019-02929-9
89. Spadaro A, Sarker A, Hogg-Bremer W, Love JS, O'Donnell N, Nelson LS, et al. Reddit discussions about buprenorphine associated precipitated withdrawal in the era of fentanyl. Clinical Toxicol-ogy. 2022;60(6):694-701. https://doi.org/10.1080/15563650.2022.2032730
90. Kogan NE, Clemente L, Liautaud P, Kaashoek J, Link NB, Nguyen AT, et al. An early warning ap-proach to monitor COVID-19 activity with multiple digital traces in near real time. Science Ad-vances. 2021;7(10):eabd6989. https://doi.org/10.1126/sciadv.abd6989
91. Bremer W, Sarker A. Recruitment and retention in mobile application-based intervention studies: a critical synopsis of challenges and opportunities. Informatics for Health and Social Care. 2023;48(2):139-52. https://doi.org/10.1080/17538157.2022.2082297
92. 92. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. Journal of Biomedical Informatics. 2015;54:202-12. https://doi.org/10.1016/j.jbi.2015.02.004
93. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using reports of symptoms and diagnoses on so-cial media to predict COVID-19 case counts in mainland China: Observational infoveillance study. Journal of medical Internet research. 2020;22(5):e19421. https://doi.org/10.2196/19421
94. Zhang H, Wheldon C, Dunn AG, Tao C, Huo J, Zhang R, et al. Mining Twitter to assess the determi-nants of health behavior toward human papillomavirus vaccination in the United States. Journal of the American Medical Informatics Association. 2020;27(2):225-35. https://doi.org/10.1093/jamia/ocz191
95. Guo Y, Ovadje A, Al-Garadi MA, Sarker A. Evaluating Large Language Models for HealthRelated Text Classification Tasks with Public Social Media Data; ArXiv. 2024. https://arxiv.org/abs/2403.19031
96. Correia RB, Wood IB, Bollen J, Rocha LM. Mining Social Media Data for Biomedical Signals and Health-Related Behavior. Annual Review of Biomedical Data Science. 2020;3(1):433-58. https://doi.org/10.1146/annurev-biodatasci-030320-040844
97. Yang K, Zhang T, Kuang Z, Xie Q, Ananiadou S. MentalLLaMA: Interpretable Mental Health Analy-sis on Social Media with Large Language Models. arXiv preprint arXiv:230913567. 2023. https://arxiv.org/abs/2309.13567
98. Lai T, Shi Y, Du Z, Wu J, Fu K, Dou Y, et al. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. arXiv preprint arXiv:230711991. 2023. https://arxiv.org/abs/2307.11991
99. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. In: Proceedings of the ACM on Interac-tive, Mobile, Wearable and Ubiquitous Technologies. 2023;8(1)31. p. 1-32. https://doi.org/10.1145/3643540
100. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, et al. Stanford Alpaca: An Instruction-following LLaMA model. GitHub; 2023. https://github.com/tatsu-lab/stanford_alpaca
101. Chen J, Wang Y. Social media use for health purposes: systematic review. Journal of medical In-ternet research. 2021;23(5):e17917. https://doi.org/10.2196/17917
102. Jiang K, Devendra V, Chavan S, Bernard GR. Detection of Day-Based Health Evidence with Pre-trained Large Language Models: A Case of COVID-19 Symptoms in Social Media Posts. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2023. p. 4208-12. Available from: https://doi.org/10.1109/BIBM58861.2023.10385580
103. Chen C, Shu K. Combating misinformation in the age of llms: Opportunities and challenges. arXiv preprint arXiv:231105656. 2023. https://arxiv.org/abs/2311.05656
104. Xiao Z, Liao QV, Zhou M, Grandison T, Li Y. Powering an AI Chatbot with Expert Sourcing to Sup-port Credible Health Information Access. In: Proceedings of the 28th International Conference on Intelligent User Interfaces; 2023. p. 2-18. https://doi.org/10.1145/3581641.3584031
105. Tran V, Matsui T. Public Opinion Mining Using Large Language Models on COVID-19 Related Tweets. In: 2023 15th International Conference on Knowledge and Systems Engineering (KSE). IEEE; 2023. p. 1-6. https://doi.org/10.1109/KSE59128.2023.10299499
106. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or genera-tive AI) in healthcare. NPJ digital medicine. 2023;6(1):120. https://doi.org/10.1038/s41746-023-00873-0

107. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. The Lancet Digital Health. 2024;6(1):e12-22. https://doi.org/10.1016/S2589-7500(23)00225-X
108. Wang X, Chen N, Chen J, Hu Y, Wang Y, Wu X, et al. Apollo: Lightweight Multilingual Medical LLMs towards Democratizing Medical AI to 6B People. arXiv preprint arXiv:240303640. 2024. https://arxiv.org/abs/2403.03640
109. Bommasani R, Klyman K, Longpre S, Kapoor S, Maslej N, Xiong B, et al. The foundation model transparency index. arXiv preprint arXiv:231012941. 2023. https://arxiv.org/abs/2310.12941
110. Lin J, Zhu L, Chen WM, Wang WC, Han S. Tiny Machine Learning: Progress and Futures [Feature]. IEEE Circuits and Systems Magazine. 2023;23(3):8-34. https://doi.org/10.1109/MCAS.2023.3302182
111. Chen Q, Du J, Hu Y, Keloth VK, Peng X, Raja K, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv preprint arXiv:230516326. 2023. https://arxiv.org/abs/2305.16326
112. Wu C, Lin W, Zhang X, Zhang Y, Wang Y, Xie W. PMC-LLaMA: Towards Building Opensource Lan-guage Models for Medicine. arXiv; 2023. ArXiv:2304.14454. http://arxiv.org/abs/2304.14454
113. French RM. Catastrophic forgetting in connectionist networks. Trends in cognitive sciences. 1999;3(4):128-35. https://doi.org/10.1016/S1364-6613(99)01294-2.
114. Gupta K, Thérien B, Ibrahim A, Richter ML, Anthony QG, Belilovsky E, et al. Continual Pre-Training of Large Language Models: How to re-warm your model? In: Workshop on Efficient Systems for Foundation Models@ ICML2023; 2023. Available from: https://openreview.net/forum?id=pg7PUJe0Tl.
115. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Communications Medicine. 2023;3(1):141. https://doi.org/10.1038/s43856-023-00370-1
116. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:231210997. 2023. https://arxiv.org/abs/2312.10997
117. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A Survey on Hallucination in Large Lan-guage Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv; 2023. ArXiv:2311.05232. http://arxiv.org/abs/2311.05232
118. Liu Y, Yao Y, Ton JF, Zhang X, Cheng RGH, Klochkov Y, et al. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv preprint arXiv:230805374. 2023. https://arxiv.org/abs/2308.05374.

Copyright