The Role of Language Models in Modern Healthcare: A Comprehensive Review

Amna Khalid
CHRISTUS Santa Rosa Hospital
New Braunfels, TX

Ayma Khalid
Riphah International University
Lahore, Pakistan

Umar Khalid
Palo Alto, CA

Abstract

The application of large language models (LLMs) in healthcare has gained significant attention due to their ability to process complex medical data and provide insights for clinical decision-making. These models have demonstrated substantial capabilities in understanding and generating natural language, which is crucial for medical documentation, diagnostics, and patient interaction. This review examines the trajectory of language models from their early stages to the current state-of-the-art LLMs, highlighting their strengths in healthcare applications and discussing challenges such as data privacy, bias, and ethical considerations. The potential of LLMs to enhance healthcare delivery is explored, alongside the necessary steps to ensure their ethical and effective integration into medical practice.

***K*eywords** Large Language Models, Healthcare, Machine Learning, Natural Language Processing, Medical AI, Ethics in AI, Clinical Decision Support

1 Introduction

Deep learning has revolutionized the way we understand human behavior, emotions, and healthcare-related challenges [1, 2, 3, 4]. In recent years, breakthroughs in clinical language processing have paved the way for transformative changes in the healthcare industry. These advancements hold great promise for the deployment of intelligent systems that can support decision-making, accelerate diagnostic workflows, and enhance the quality of patient care. Such systems have the potential to assist healthcare professionals as they navigate the growing body of medical knowledge, interpret complex patient records, and craft individualized treatment plans. The promise of these systems has generated significant excitement within the healthcare community [5, 6, 7].

The power of large language models (LLMs) lies in their ability to analyze vast amounts of medical literature, patient data, and the rapidly growing body of clinical research. Healthcare data [8, 9] is inherently intricate, heterogeneous, and extensive. LLMs function as critical tools that help alleviate information overload for healthcare professionals. By automating the processing of medical texts, extracting key insights, and applying the knowledge, LLMs have the potential to drive significant research breakthroughs and improve patient care, contributing meaningfully to the evolution of the medical field.

The excitement surrounding LLMs is largely driven by the impressive capabilities of advanced models like OpenAI's GPT-3.5, GPT-4 [10, 11], and Google's Bard. These models have shown remarkable proficiency across a broad range of natural language understanding tasks, underscoring their pivotal role in healthcare applications. With their ability to comprehend and generate human-like text, these models are set to have a transformative impact on healthcare, where accurate communication and information management are paramount [12].

Natural language processing (NLP) has undergone significant advancements, with each milestone building on the strengths and limitations of previous approaches. Early developments, such as recurrent neural networks (RNNs), laid the groundwork for contextual understanding in NLP tasks. However, their limitations in handling long-range dependencies became clear, necessitating new approaches in the field.

The turning point came with the introduction of the Transformer architecture, which effectively addressed the challenge of capturing distant relationships between words. This innovation was crucial for the development of more advanced NLP models. The advent of sophisticated language models such as Llama 2 [13] and GPT-4, both of which benefit from extensive training datasets, has propelled NLP to new heights, allowing for deeper understanding and near-human-level text generation.

Within healthcare, specialized versions of models like BERT, including BioBERT and ClinicalBERT [14, 15], were developed to address the unique challenges of clinical language, such as medical terminology, ambiguity, and variability in usage. However, the use of LLMs in the highly sensitive healthcare sector requires careful consideration of privacy, security, and ethics. Patient data must be rigorously protected, and models must be designed to avoid perpetuating biases or causing harm. Despite these challenges, the potential for LLMs to improve healthcare outcomes and drive innovation remains a key focus of ongoing research and development.

This review serves as a comprehensive guide for medical researchers and healthcare professionals aiming to optimize the use of LLMs in their practices. It provides a detailed exploration of LLM technologies, their applications in healthcare, and critical discussions on fairness, bias, privacy, transparency, and ethical considerations. By addressing these aspects, this review highlights the importance of integrating LLMs into healthcare in a responsible, equitable, and effective manner to maximize benefits for both patients and providers.

The paper is organized into the following sections:

- **Section 2** introduces the fundamental architecture of LLMs, including key components such as Transformers, foundational models, and their multi-modal capabilities.
- **Section 3** explores the application of LLMs in healthcare, detailing their various use cases and the performance metrics used to evaluate them in clinical environments.
- **Section 4** delves into the challenges that LLMs face in healthcare, focusing on issues such as explainability, security, bias, and ethical concerns.
- Finally, the paper concludes with a summary of the findings, discussing the transformative potential of LLMs while addressing the need for careful implementation to mitigate limitations and ethical challenges.

2 Overview of Large Language Models

Large language models (LLMs) have rapidly advanced due to their ability to understand and generate human-like text across a variety of natural language processing (NLP) tasks [16, 10]. These models are distinguished by their extensive number of parameters, pre-training on vast text datasets, and subsequent fine-tuning for specific tasks [17, 18, 13]. In this section, we examine the core architecture of LLMs, highlight key examples, and explore pre-training methodologies as well as the role of transfer learning [19].

LLMs leverage the Transformer architecture, which excels in capturing long-range dependencies within text [20]. The self-attention mechanism inherent to this architecture enables models to focus on different parts of the input text based on their relevance, improving the handling of complex linguistic relationships.

2.1 Transformers and Their Role in Language Models

A hallmark of LLMs is their scale [21, 22], pre-training on immense text corpora [23, 13], and the fine-tuning process tailored to particular tasks [24]. These models, composed of billions of parameters, are designed to recognize intricate patterns in language data. After undergoing broad pre-training, they are refined using smaller, task-specific datasets, resulting in enhanced performance across a variety of NLP applications.

The introduction of the Transformer framework revolutionized the field by addressing the limitations of earlier architectures like recurrent neural networks (RNNs) [20]. This evolution led to the development of powerful models like GPT-4 [11] and Llama 2 [13], significantly improving natural language understanding and generation.

2.2 Multi-Modal Language Models: Expanding Capabilities

A significant progression in AI is the rise of multi-modal language models (MLLMs), which integrate data from multiple sources, such as text, images, and audio. These models, such as BLIP-2 [25], extend the traditional capabilities of LLMs by incorporating multiple modalities, allowing for more versatile and robust outputs [26]. MLLMs enable tasks such as visual question answering (VQA) and cross-modal content generation, opening up new possibilities for real-world applications.

Table 1: Summary of Multi-Modal Language Models

Model	Year	Capabilities	Applications
BLIP-2 [25]	2023	Image-text integration using Qformer	Visual question answering, image-text retrieval
Visual ChatGPT [26]	2023	Text and image interaction via GPT	Complex queries requiring visual inputs
MoVA [27]	2024	Mixture of experts for image and text	Multi-modal content generation and analysis

Table 2: Overview of Large Language Models in Healthcare

Model	Year	Use Case	Institution	Source Code
BioMistral [30]	2024	Medical Question Answering	Avignon Université, Nantes Université	model
Med-PaLM 2 [31]	2023	Medical Question Answering	Google Research, DeepMind	
Radiology-Llama2 [32]	2023	Radiology Imaging Analysis	University of Georgia	
DeID-GPT [33]	2023	Data De-identification	University of Georgia	code
Med-HALT [34]	2023	Hallucination Detection	Saama AI Research	code
ChatCAD [35]	2023	Computer-Aided Diagnosis	ShanghaiTech University	code
BioGPT [36]	2023	Classification, Relation Extraction, Question Answering	Microsoft Research	code
GatorTron [37]	2022	Medical Textual Similarity, Inference, Question Answering	University of Florida	code

2.3 Applications of Large Language Models in Healthcare

LLMs have also become prominent in healthcare, where they support tasks such as medical diagnostics, patient care, and drug discovery [28, 29]. Tailored models like BioBERT [14] and ClinicalBERT [15] are designed to handle the specialized language found in medical records and research. Newer models, including GPT-4 and Google's Bard, are setting new benchmarks in medical question answering and related healthcare applications [6].

2.4 Real-World Healthcare Applications of Large Language Models

LLMs have been widely adopted across various healthcare functions, with applications continuing to expand rapidly. These models assist in clinical decision-making, analysis of medical records, and improving patient interactions [38]. The vast capability of LLMs to process medical data offers benefits in areas such as diagnostics, administrative efficiency, and overall healthcare delivery [39, 40].

- **Medical Diagnostics:** LLMs can help physicians diagnose illnesses by analyzing patient data, including symptoms and medical histories, to identify potential health conditions [41].
- **Patient Care:** Through personalized recommendations and ongoing patient monitoring, LLMs improve the quality of patient care by providing real-time insights [42].
- **Clinical Decision Support:** LLMs offer healthcare professionals evidence-based recommendations, enhancing clinical decision-making and treatment strategies [43].
- **Medical Literature Review:** By summarizing large volumes of medical literature, LLMs help healthcare professionals stay current with new developments and best practices [44].
- **Drug Discovery:** LLMs facilitate drug discovery by analyzing molecular data to identify potential compounds for new drugs [28, 45].

Table 3: Evaluation Metrics for LLMs in Healthcare Applications

Metric	Task	Description	Key Results
Perplexity	Language Generation	Measures model uncertainty	Lower perplexity indicates better language generation performance
BLEU	Translation	Evaluates overlap between generated and reference text	ClinicalGPT achieved a BLEU score of 13.9 [48]
ROUGE	Summarization	Assesses recall of generated summaries	BioMedLM attained a ROUGE-L score of 24.85 [49]
F1 Score	Classification	Combines precision and recall for a balanced metric	GatorTron obtained an F1 score of 0.9627 for medical relation extraction [37]

Table 4: Benchmark Comparison of Large Language Models

Model	MMLU Score	HumanEval (Coding)	Release Date
GPT-4 Turbo	86.4	85.4	April 2024
Claude 3.5	88.7	92.0	June 2024
Llama 3	86.1	81.7	March 2024
Gemini Ultra	83.7	74.3	December 2023

- **Virtual Health Assistants:** LLMs serve as the backbone for healthcare chatbots that provide continuous health monitoring and medical advice [46].
- **Radiology and Imaging:** Multi-modal LLMs assist radiologists by analyzing imaging data and improving diagnostic precision [47].
- **Automated Report Generation:** LLMs automate the generation of medical reports from diagnostic images, speeding up workflows in radiology and pathology [35].

2.5 Performance Metrics and Model Comparisons

Benchmarking LLM performance is crucial for assessing their effectiveness across different healthcare tasks. Commonly used benchmarks, such as MMLU (Massive Multitask Language Understanding) and HumanEval, evaluate LLMs on various tasks, including problem-solving and code generation [50, 51]. Table 4 presents a comparison of several state-of-the-art models based on these benchmarks.

3 Challenges and Future Directions

The incorporation of large language models (LLMs) in healthcare is not without obstacles. These hurdles include the need for greater transparency in model decisions, ensuring data privacy and security for sensitive patient information, addressing biases to guarantee fairness, preventing the generation of false or misleading outputs, and establishing regulatory frameworks for ethical AI use in medical contexts. Overcoming these challenges is vital for fully harnessing LLMs' potential to improve healthcare while maintaining ethical and legal standards.

3.1 Improving Model Transparency and Interpretability

One significant challenge when applying LLMs in healthcare is their lack of interpretability. These models often function as "black boxes," making it difficult for healthcare providers to understand how specific recommendations or predictions are generated. This lack of clarity can hinder adoption, as medical professionals require transparent decision-making processes to ensure accuracy and trust. In healthcare, where every decision must be well-founded, the opaque nature of LLMs is particularly problematic. To address this, efforts are underway to develop more interpretable models that offer insight into their decision-making processes, fostering trust in AI-generated recommendations [52, 53]. Enhancing transparency and interpretability remains a key research focus in healthcare AI [54, 55, 56].

3.2 Data Privacy and Security Risks

When applied in healthcare settings, LLMs handle vast amounts of sensitive information, including personally identifiable data. Ensuring this data is processed and stored securely, in compliance with privacy regulations, is a significant challenge. One concern is the unintentional exposure of personal health information (PHI) during the training process, which could lead to privacy violations. Furthermore, the ability of LLMs to infer sensitive information from anonymized data presents additional privacy risks [57]. To mitigate these threats, it is essential to implement robust anonymization techniques, secure data storage, and compliance with ethical guidelines, ensuring that patient data remains protected throughout the use of LLMs in healthcare [58].

3.3 Ensuring Fairness and Reducing Bias

LLMs can inherit biases from the data they are trained on, particularly if the datasets include unequal representations of demographic groups or healthcare outcomes. These biases can lead to disparities in medical recommendations and outcomes, which can be harmful in clinical settings. Researchers must develop strategies to identify, reduce, and prevent biases within these models, ensuring that LLMs contribute to equitable healthcare solutions. Ongoing audits and evaluations are critical for identifying and mitigating biases in both training data and model outputs [58]. Collaboration between domain experts, data scientists, and ethicists can foster the development of fair and unbiased AI in healthcare.

3.4 Preventing Hallucinations in Medical AI

LLMs sometimes generate false or misleading information—commonly referred to as hallucinations—which can be particularly dangerous in healthcare applications where accuracy is critical. These models may produce plausible-sounding, but factually incorrect, content without providing traceable sources [59]. Healthcare professionals must be cautious when using LLMs, validating AI-generated content to avoid the risks associated with incorrect medical guidance. Current research is focused on addressing these hallucination challenges, with benchmarks like Med-HALT being developed to evaluate how well models perform in medical reasoning and information retrieval [34].

3.5 Legal, Ethical, and Regulatory Frameworks

The use of LLMs in healthcare also raises significant legal and ethical questions. Issues such as the generation of sensitive or distressing medical content, or the potential for spreading misinformation, necessitate strict regulatory oversight. Furthermore, there are concerns about plagiarism, impersonation, and the overall integrity of LLM-generated content. Regulatory frameworks, such as the EU's AI Act and the U.S. HIPAA, provide essential guidelines for the safe and responsible deployment of AI in healthcare [57, 60]. These laws ensure patient data protection and set ethical standards for the use of AI technologies in sensitive environments, fostering trust and accountability in AI-powered healthcare.

4 Closing Remarks

The adoption of large language models in healthcare presents substantial opportunities for enhancing medical decision-making and information retrieval. These models, equipped with advanced capabilities, have the potential to improve workflows and patient outcomes across various healthcare applications. However, realizing their full potential requires overcoming key challenges such as ensuring model transparency, protecting sensitive data, reducing biases, and preventing erroneous outputs. As researchers and practitioners continue to collaborate, the focus must remain on developing ethical, trustworthy, and fair AI systems that meet the rigorous standards of healthcare. Continued innovation, combined with careful consideration of ethical and regulatory concerns, will shape the future of LLMs in medical practice.

Table 5: Overview of Challenges and Mitigation Strategies for LLMs in Healthcare

Challenge	**Impact**	**Proposed Solution**
Transparency	Lack of understanding in AI-generated decisions	Develop interpretable models and provide decision explanations
Data Security	Risk of exposing sensitive patient information	Use advanced anonymization and secure data storage protocols
Bias	Perpetuation of unfair treatment outcomes	Conduct regular bias audits and collaborate with domain experts
Hallucinations	Creation of inaccurate or misleading content	Implement rigorous validation and specialized benchmarks like Med-HALT
Ethical and Legal Concerns	Risk of misuse and data breaches	Comply with regulations such as HIPAA and the AI Act, and ensure ethical use of AI

References

[1] Henglin Shi, Wei Peng, Haoyu Chen, Xin Liu, and Guoying Zhao. Multiscale 3d-shift graph convolution network for emotion recognition from human actions. *IEEE Intelligent Systems*, 37(4):103–110, 2022.

[2] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11185–11195, 2023.

[3] Yante Li, Wei Peng, and Guoying Zhao. Micro-expression action unit detection with dual-view attentive similarity-preserving knowledge distillation. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[4] Xiaopeng Hong, Wei Peng, Mehrtash Harandi, Ziheng Zhou, Matti Pietikäinen, and Guoying Zhao. Characterizing subtle facial movements via riemannian manifold. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–24, 2019.

[5] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*, 2023.

[6] Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*, 2023.

[7] Ping Yu, Hua Xu, Xia Hu, and Chao Deng. Leveraging generative ai and large language models: a comprehensive roadmap for healthcare integration. In *Healthcare*, volume 11, page 2776. MDPI, 2023.

[8] Wei Peng, Li Feng, Guoying Zhao, and Fang Liu. Learning optimal k-space acquisition and reconstruction using physics-informed neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20794–20803, 2022.

[9] Wei Peng, Ehsan Adeli, Tomas Bosschieter, Sang Hyun Park, Qingyu Zhao, and Kilian M Pohl. Generating realistic brain mris via a conditional diffusion probabilistic model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–24. Springer, 2023.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[11] OpenAI. Gpt-4 technical report, 2023.

[12] Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, et al. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint arXiv:2304.06488*, 2023.

[13] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[14] 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, 2020.

[15] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

[16] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[17] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[18] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[19] 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.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[21] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

[22] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.

[23] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-trained language models and their applications. *Engineering*, 2022.

[24] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[26] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

[27] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.

[28] Zhichao Liu, Ruth A Roberts, Madhu Lal-Nag, Xi Chen, Ruili Huang, and Weida Tong. Ai-based language models powering drug discovery and development. *Drug Discovery Today*, 26(11):2593–2607, 2021.

[29] Tanmoy Tapos Datta, Pintu Chandra Shill, and Zabir Al Nazi. Bert-d2: Drug-drug interaction extraction using bert. In *2022 International Conference for Advancement in Technology (ICONAT)*, pages 1–6. IEEE, 2022.

[30] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[31] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[32] Zhengliang Liu, Yiwei Li, Peng Shu, Aoxiao Zhong, Longtao Yang, Chao Ju, Zihao Wu, Chong Ma, Jie Luo, Cheng Chen, et al. Radiology-llama2: Best-in-class large language model for radiology. *arXiv preprint arXiv:2309.06419*, 2023.

[33] Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023.

[34] Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

[35] Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. Chatcad+: Towards a universal and reliable interactive cad using llms. *arXiv preprint arXiv:2305.15964*, 2023.

[36] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

[37] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, et al. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *arXiv preprint arXiv:2203.03540*, 2022.

[38] Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*, 2022.

[39] Hongyang Li, Richard C Gerkin, Alyssa Bakke, Raquel Norel, Guillermo Cecchi, Christophe Laudamiel, Masha Y Niv, Kathrin Ohla, John E Hayes, Valentina Parma, et al. Text-based predictions of covid-19 diagnosis from self-reported chemosensory descriptions. *Communications Medicine*, 3(1):104, 2023.

[40] Felix Agbavor and Hualou Liang. Predicting dementia from spontaneous speech using large language models. *PLOS Digital Health*, 1(12):e0000168, 2022.

[41] Zekai Chen, Mariann Micsinai Balan, and Kevin Brown. Boosting transformers and language models for clinical prediction in immunotherapy. *arXiv preprint arXiv:2302.12692*, 2023.

[42] Stephen R Ali, Thomas D Dobbs, Hayley A Hutchings, and Iain S Whitaker. Using chatgpt to write patient clinic letters. *The Lancet Digital Health*, 5(4):e179–e181, 2023.

[43] Rubeta N Matin, Eleni Linos, and Neil Rajan. Leveraging large language models in dermatology, 2023.

[44] Malik Sallam. The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv*, pages 2023–02, 2023.

[45] Gökçe Uludoğan, Elif Ozkirimli, Kutlu O Ulgen, Nilgün Karalı, and Arzucan Özgür. Exploiting pretrained biochemical language models for targeted drug design. *Bioinformatics*, 38(Supplement_2):ii155–ii161, 2022.

[46] Desirée Bill and Theodor Eriksson. Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application, 2023.

[47] Lei Ma, Jincong Han, Zhaoxin Wang, and Dian Zhang. Cephgpt-4: An interactive multimodal cephalometric measurement and diagnostic system with visual large language model. *arXiv preprint arXiv:2307.07518*, 2023.

[48] Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023.

[49] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*, 2023.

[50] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[51] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[52] Hazrat Ali, Junaid Qadir, Tanvir Alam, Mowafa Househ, and Zubair Shah. Chatgpt and large language models (llms) in healthcare: Opportunities and risks. 2023.

[53] Sandeep Reddy. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, page 101304, 2023.

[54] Giovanni Briganti. A clinician's guide to large language models. *Future Medicine AI*, (0):FMAI, 2023.

[55] Aleksa Bisercic, Mladen Nikolic, Mihaela van der Schaar, Boris Delibasic, Pietro Lio, and Andrija Petrovic. Interpretable medical diagnostics with structured data extraction by large language models. *arXiv preprint arXiv:2306.05052*, 2023.

[56] Yan Jiang, Ruihong Qiu, Yi Zhang, and Peng-Fei Zhang. Balanced and explainable social media analysis for public health with large language models. *arXiv preprint arXiv:2309.05951*, 2023.

[57] Jesutofunmi A Omiye, Haiwen Gui, Shawheen J Rezaei, James Zou, and Roxana Daneshjou. Large language models in medicine: the potentials and pitfalls. *arXiv preprint arXiv:2309.00087*, 2023.

[58] Surendrabikram Thapa and Surabhi Adhikari. Chatgpt, bard, and large language models for biomedical research: Opportunities and pitfalls. *Annals of Biomedical Engineering*, pages 1–5, 2023.

[59] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *arXiv preprint arXiv:2306.10070*, 2023.

[60] Claudio Novelli, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. Generative ai in eu law: liability, privacy, intellectual property, and cybersecurity. *arXiv preprint arXiv:2401.07348*, 2024.