Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 8, S. 46-54, 9, 2024

Araştırma Makalesi

IJANSER

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 8, pp. 46-54, 9, 2024

Research Article

https://as-proceeding.com/index.php/ijanser
ISSN:2980-0811

Application of Large Language Models (LLMs) in the Field of Healthcare

Jandoubi Aymen*, El Hamdi Ridha and Njah Mohamed

Advanced Technologies Medical & Signals.
National Engineering School of Sfax, Tunisia.
Digital Research Center of Sfax, Technopole of Sfax, Tunisia.

*Aymenjendoubi13@gmail,com

(Received: 25 September 2024, Accepted: 02 October 2024)

(6th International Conference on Applied Engineering and Natural Sciences ICAENS 2024, 25-26 September 2024)

ATIF/REFERENCE: Aymen, J., Ridha, E. H. & Mohamed, N. (2024). Application of Large Language Models (LLMs) in the Field of Healthcare, *International Journal of Advanced Natural Sciences and Engineering Researches*, 8(9), 46-54.

Abstract **–** Large language models (LLMs) are revolutionizing healthcare by integrating advanced natural language processing and machine learning technologies. This proposal outlines a survey to explore LLMs' roles in healthcare, focusing on their development, performance, practical applications, and challenges. The survey will examine how LLMs can enhance medical education, clinical decision-making, and manage complex medical data for personalized care. Additionally, it will assess LLMs' impact on medical workflows, research, and diagnostics, addressing reliability, safety, and ethical considerations. This survey aims to provide insights into LLMs' transformative potential and guide future research and innovation.

Keywords – LLMs, NLP, Healthcare, Clinical decision-making, Medical education, Electronic health records, Medical imaging, Personalized care.

I. Introduction

Large Language Models (LLMs) such as GPT-4 signify a significant leap forward in artificial intelligence, particularly in the realm of natural language processing. These systems have the remarkable ability to learn, comprehend, and produce human language in intricate and nuanced manners, thereby paving the way for innovation across various industries, including healthcare.

Within the medical domain, LLMs hold immense transformative potential. They can sift through extensive medical literature, aid in clinical decision-making, manage patient records efficiently, and even contribute to medical training and education. Such capabilities hint at a future where LLMs could become pivotal in healthcare delivery, providing assistance to both healthcare professionals and patients alike.

This article aims to accomplish two main objectives. Firstly, it aims to evaluate the current utilization of LLMs in healthcare, including their current applications, recent advancements, and the feedback from healthcare practitioners and patients regarding their experiences with these technologies. Secondly, it endeavors to delve into the future prospects of LLMs in the field of medicine, focusing on potential advancements, hurdles to overcome, and the ethical and practical considerations associated with the growing integration of LLMs into healthcare practices. In essence, this article aims to offer a comprehensive overview of both the present and future impact of Large Language Models on the healthcare sector.

II. STRATEGIES

The incorporation of Large Language Models (LLMs) into the medical industry has ignited significant research attention. This section delineates the fundamental approaches employed in the development of medical LLMs and offers an overview of the broader development of LLMs. Typically, the development of medical LLMs revolves around three primary methods: starting from scratch with pre-training, refining pre-existing LLMs through fine-tuning, or employing direct prompts to tailor general LLMs for medical use (see Figure 1).

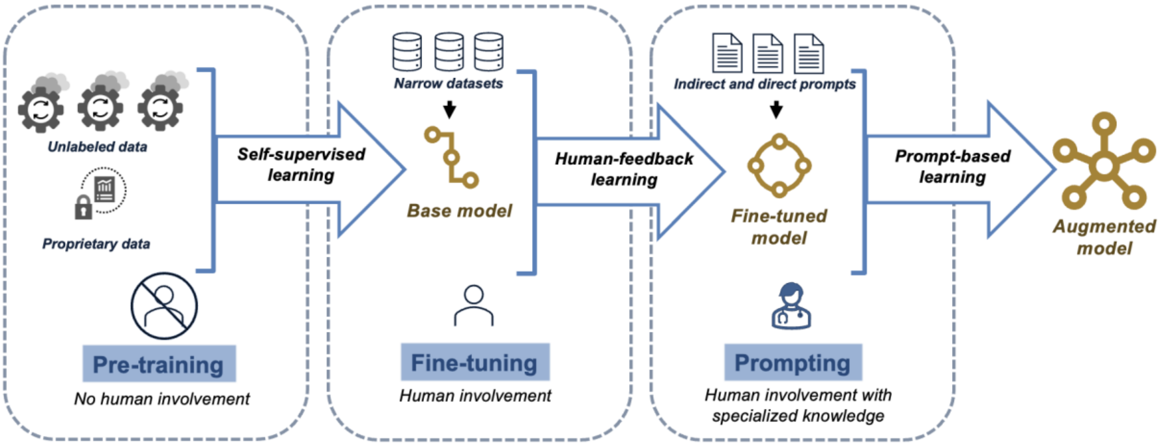

Figure 1: Overview of LLM training process.

A. *Pre-training*

Pre-training of medical LLMs involves training on various medical texts, such as electronic health records (EHRs) [1], clinical notes [2], DNA sequences [3], and medical literature from sources like PubMed [5] and MIMIC-III [6]. Models like PubMedBERT [7] and ClinicalBERT [4] are developed through this method. Objectives during pre-training include masked language modeling and next sentence prediction, focusing on the medical domain. These models are then applied to downstream tasks like question answering (QA) [11] and named entity recognition (NER), which are essential for medical research and diagnostics.

The pretrained medical LLMs are summarized in Table1.

Table 1: Overview of current medical-domain large lan- guage models, focusing on their development through pre-training, including the scale of their parameters, the datasets utilized, and their sources of data.

Models	Parameters	Data Scale	Data Source
BioBERT [9]	110M	18B tokens	PubMed [5]
PubMedBERT [7]	110M/340M	3.2B tokens	PubMed [5]
SciBERT [40]	110M	3.17B tokens	Literature [45]
ClinicalBERT [4]	110M	112k clinical notes	MIMIC-III [6]
BlueBERT [8]	110M/340M	4.5B tokens	MIMIC-III [5]
BioCPT [41]	330M	255M articles	PubMed [5]
BioGPT [42]	1.5B	15M articles	PubMed [5]
BioMedLM [43]	2.7B	110GB	PubMed [5]
OphGLM [44]	6.2B	20k dialogues	MedDialog [46]

B. ***Fine-tuning***

Given the high computational costs of training LLMs from scratch, fine-tuning existing models with medical data is a popular approach. Techniques like Supervised Fine-Tuning (SFT) and Instruction Fine-Tuning (IFT) refine these models' understanding of medical contexts. Models such as DoctorGLM [14] have been fine-tuned using physician-patient dialogue, demonstrating improved medical task performance. Parameter-efficient methods like Low-Rank Adaptation (LoRA) reduce the computational demands of fine-tuning, making it a cost-effective strategy for developing specialized medical LLMs.

The fine-tuned medical LLMs are outlined in Table 2.

Table 2: Overview of current medical-domain large lan- guage models, focusing on their development through fine- tuning, including the scale of their parameters, the datasets utilized, and their sources of data.

Models	Parameters	Data Scale	Data Source
DoctorGLM [14]	6.2B	323MB dialogues	CMD [48]
ClinicalGPT [49]	7B	100k dialogues	MedDialog [46]
Qilin-Med [17]	7B	3GB	ChiMed [17]
ChatDoctor [15]	7B	110k dialogues	iCliniq [50]
BenTsao [16]	7B	8k instructions	CMeKG-8K [51]
MedAlpaca [16]	7B/13B	160k medical QA	Medical Meadow [16]

C. ***Prompting***

Prompting strategies, including zero-shot and chain-of-thought prompting, offer a way to align LLMs with medical tasks without extensive retraining or dataset curation. These methods prompt LLMs to perform specific medical tasks or reasoning, with minimal computational overhead. MedPaLM [12] and other models employ these techniques for tasks such as medical question-answering, achieving high performance with reduced resource requirements. Prompt tuning, involving trainable continuous vectors, represents a flexible and efficient approach to adapting LLMs for medical applications.

The prompted medical LLMs are summarized in Table 3.

Table 3: Overview of current medical-domain large lan- guage models, focusing on their development through prompting, including the scale of their parameters, the datasets utilized, and their sources of data.

Models	Parameters	Data Scale	Data Source
DeID-GPT [20]	ChatGPT/GPT-4	Chain-of-Thought	-
ChatCAD [23]	ChatGPT	Zero-shot Prompting	-
Dr. Knows [22]	ChatGPT	Zero-shot Prompting	UMLS [13]
MedPaLM [12]	PaLM (540B)	40 instructions	MultiMedQA [13]
MedPrompt [21]	GPT-4	Few-shot Prompting	-

III. Clinical Applications

This section delves into the utilization of Large Language Models (LLMs) in clinical settings. For every subsection, we begin by presenting the application, followed by an exploration of how LLMs are employed to fulfill these tasks. We then address the challenges faced by LLMs within these specific contexts and conclude with a look at potential future developments and directions for LLMs in these areas of application.

1. *Diagnosis Enhancement*

LLMs contribute to medical diagnostics by integrating objective data and patient symptoms to improve disease identification. Timely, accurate diagnoses are critical, especially for conditions like breast cancer, where early detection significantly impacts survival rates. LLMs, such as Dr. Knows, utilize graph models and the Unified Medical Language System to prioritize diagnoses, showing an 8-18% improvement in accuracy [22]. However, their reliance on textual inputs limits their ability to process medical images directly, necessitating complementary tools like ChatCAD for image analysis [23]. Despite advances in vision-capable LLMs, challenges in privacy, accountability, and bias remain [28].

2. *Streamlining Clinical Reporting*

Clinical reporting, crucial yet cumbersome, benefits from LLMs by reducing errors and workloads [23]. These models summarize diagnostic information, aiding in the creation of coherent reports from images and physician inputs, potentially improving diagnostic performance by 16.42% [23]. Despite their utility, LLM-generated reports may suffer from inaccuracies or lack the nuanced understanding of human-written documents [28].

3. *Advancing Medical Education*

LLMs find significant applications in medical education, from enhancing exam preparation to acting as interactive learning aids. By generating diverse educational content, LLMs expose students to a wider range of scenarios, promoting critical thinking and adaptability [28]. They also demystify medical jargon for the public, improving health literacy [28]. Nevertheless, ethical concerns, biases, and the risk of misinformation pose challenges to their educational use [28].

IV. Challenges

Implementing Large Language Models (LLMs) in the medical field encounters challenges such as extensive computational demands and privacy considerations. Furthermore, LLMs may generate erroneous information, referred to as "hallucination" [29], and display data bias, posing ethical dilemmas [30].

However, despite these hurdles, the outlook for LLMs in healthcare appears promising. Ongoing research endeavors are focused on addressing these challenges to facilitate broader adoption, consequently enhancing personalized medicine and the quality of patient care.

1. *Mitigating Hallucination*

LLMs can generate misleading information (hallucination), including intrinsic errors, such as false mathematical outputs [29], and extrinsic errors, like fabricating references. In healthcare, such inaccuracies can lead to detrimental outcomes like misdiagnoses. Addressing LLM hallucinations involves strategies like training adjustments to minimize errors [31], inference enhancements for better reliability [32], and using external facts for validation [33].

2. *Improving Evaluation Methods*

The advancement of LLMs outpaces current benchmarks, making it challenging to assess their effectiveness in healthcare. Current benchmarks focus on question-answering but miss evaluating essential attributes like trustworthiness [35]. Proposals for more relevant benchmarks, such as HealthSearchQA [12], aim to align evaluations more closely with human needs in healthcare. Developing benchmarks that measure medical and LLM-specific metrics is crucial for accurate performance evaluation.

3. *Incorporating New Knowledge*

Updating LLMs with the latest medical knowledge is hampered by the difficulty in removing outdated information and the challenge of timely updates. Solutions include model editing for direct knowledge updates and retrieval-augmented generation for leveraging external knowledge sources, such as updating external memory for real-time information relevance [36]. These strategies aim to enhance LLMs' accuracy and timeliness in medical knowledge application.

4. *Data Restrictions within the Domain*

Presently, datasets within the medical domain are notably smaller in comparison to those utilized for training general-purpose LLMs. Despite the vast expanse of medical knowledge, existing datasets are constrained and do not encompass the entirety of this domain. Consequently, while LLMs demonstrate exceptional performance on standardized benchmarks with extensive data coverage, they often struggle when applied to real-world tasks such as differential diagnosis and personalized treatment planning.

To address these challenges, potential solutions must be explored. Despite the abundance of medical and health data, accessing them typically involves navigating through extensive ethical, legal, and privacy protocols. Additionally, these datasets frequently lack labeling, and approaches to utilize them, such as human labeling and unsupervised learning, encounter obstacles due to limited human expert resources and narrow margins of error. Current state-of-the-art methodologies lean towards fine-tuning on smaller openly accessible datasets to enhance the models' domain-specific performance. Another avenue involves generating high-quality synthetic datasets using LLMs to expand knowledge coverage. However, studies have indicated that training on generated datasets may lead to model forgetting. Hence, future research is essential to validate the efficacy of employing synthetic data for LLMs in the medical field.

V. FUTURE DIRECTIONS

Although Large Language Models (LLMs) have already made a substantial impact on people's lives through applications such as chatbots and search engines, their integration into medicine is still in its early stages. There are numerous untapped opportunities for researchers and practitioners to enhance the role of medical LLMs in serving the public more effectively. These opportunities encompass the introduction of new benchmarks, fostering interdisciplinary collaborations, the development of multimodal LLMs, and the application of LLMs in less explored areas of medicine.

1. *Developing Comprehensive Benchmarks*

The need for benchmarks that accurately assess LLMs in clinical contexts is evident, as current measures often overlook the multifaceted skills required in healthcare. These new benchmarks should not only test medical knowledge accuracy but also evaluate how LLMs source information, adapt to new medical insights, and communicate uncertainties [12]. Furthermore, they must consider ethical dimensions like fairness and equity, challenging due to their qualitative nature [12].

Special attention is warranted for fields like rehabilitation and sports medicine, where LLMs can significantly contribute to combating global health issues like physical inactivity. With over a quarter of the world's adult population affected, LLMs could help disseminate physical activity knowledge and customize programs, especially in under-resourced areas [37].

2. Exploring Multimodal LLMs in Medicine

Multimodal LLMs (MLLMs) extend LLM capabilities to include visual, audio, and time-series data analysis, presenting new possibilities in medical diagnostics and treatment planning [38]. Innovations like MedPaLM M and Visual Med-Alpaca illustrate MLLMs' potential in interpreting medical images and integrating diverse data types for holistic patient assessments [39]. Despite their promise, challenges remain in data privacy, quality assurance, and the enhancement of MLLMs' perception and reasoning abilities [38].

3. Innovating with Medical Agents

The concept of LLM-powered medical agents, specialized in roles like radiology or pathology, offers a novel approach to disease diagnosis and treatment. By simulating the expertise of different medical specialists, these agents could collaborate for comprehensive patient evaluations, improving diagnostic accuracy and efficiency [47]. Critical to their development are considerations around data privacy, the validation of diagnostic interpretations, and the ethical implications of AI in decision-making roles.

In summary, the future of medical LLMs involves tackling complex challenges through innovative approaches, ensuring that these technologies contribute positively to healthcare outcomes.

VI. CONCLUSION

This article delves deeply into the advancement and potential uses of large language models (LLMs) within the medical sphere. It meticulously examines how these models have evolved, considering factors such as their architecture, size, and training techniques. Despite demonstrating promising results in standardized tests, there exists a noticeable gap between these outcomes and their actual efficacy within clinical settings.

Moreover, the paper explores the transformative possibilities of LLMs across various healthcare domains, including diagnostics, clinical note generation, medical education, and more. However, it also acknowledges the hurdles that need to be addressed, such as the potential for generating inaccurate information, the lack of transparency in decision-making processes, data scarcity, and the limitations of current evaluation methods.

Given that this field is still evolving, significant research efforts are needed to establish more pertinent evaluation criteria that prioritize reliability, safety, and fairness. It also emphasizes the importance of fostering closer collaborations between the medical and artificial intelligence communities. Additionally, there's a spotlight on the potential of multimodal LLMs, which can utilize diverse data types like visual and auditory information, as well as expanding LLM applications to cover a broader spectrum of medical specialties.

Ultimately, the article underscores the extensive role of LLMs in the medical domain, advocating for continuous exploration and innovation in this interdisciplinary field. While acknowledging the potential of LLMs to advance clinical care and medical research, it emphasizes the importance of responsible and effective implementation, stressing the need for ongoing collaborative efforts involving clinicians to ensure that these technologies benefit society equitably.

REFERENCES

[1] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, et al. (2023). A study of generative large language model for medical research and healthcare. *arXiv preprint arXiv:2305.13523*..

[2] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, et al. (2022). A large language model for electronic health records. *NPJ Digital Medicine*, 5(1):194.

[3] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

[4] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. (2019). Publicly available clinical BERT embeddings.

[5] National Institutes of Health. (2022). PubMed Corpora. Retrieved from PubMed.

[6] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

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

[8] Yifan Peng, Shankai Yan, and Zhiyong Lu. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 58–65).

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

[10] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. (2023). Meditron 70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

[11] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. (2021). What disease does this patient have? A large scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

[12] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

[13] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. (2023). Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

[14] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. (2023). DoctorGLM: Fine-tuning your Chinese doctor is not a Herculean task. *arXiv preprint arXiv:2304.01097*.

[15] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. (2023). ChatDoctor: A medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge.

[16] Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. (2023). Medalpaca–an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*.

[17] Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. (2023). Qilin-Med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

[18] Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. (2023). Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

[19] Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. (2023). Huatuo: Tuning LLaMA model with Chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

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

[21] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. (2023). Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*.

[22] Yanjun Gao, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M. Churpek, and Majid Afshar. (2023). Leveraging a medical knowledge graph into large language models for diagnosis prediction. *arXiv e-prints*, pages arXiv–2308.

[23] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. (2023). ChatCAD: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*.

[24] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, et al. (2023). Clinical text summarization: Adapting large language models can outperform human experts. *arXiv preprint arXiv:2309.07430*.

[25] Paul K. Drain, Aron Primack, D. Dan Hunt, Wafaie W. Fawzi, King K. Holmes, and Pierce Gardner. (2007). Global health in medical education: A call for more training and opportunities. *Academic Medicine*, 82(3):226–230.

[26] Tim Swanwick. (2018). Understanding medical education. *Understanding Medical Education: Evidence, Theory, and Practice*, pages 1–6.

[27] Mohammad H. Rajab, Abdalla M. Gazarin, and Abdullah A. Alazzeh. (2022). Applications of artificial intelligence in medical education: A review. *Journal of Education and Practice*, 13(1):1–12.

[28] Mert Karabacak, Burak Berksu Ozkara, Konstantinos Margetis, Max Wintermark, Sotirios Bisdas, et al. The advent of generative language models in medical education. JMIR Medical Education, 9(1):e48163, 2023.

[29] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12):1–38, 2023.

[30] Malik Sallam. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In Healthcare, page 887. MDPI, 2023.

[31] Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. arXiv preprint arXiv:2306.00186, 2023.

[32] Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896, 2023.

[33] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567, 2021.

[34] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Conference on Health, Inference, and Learning, pages 248–260. PMLR, 2022.

[35] Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. Faithful ai in medicine: A systematic review with large language models and beyond. medRxiv.

[36] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrievalaugmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474, 2020.

[37] World Health Organization. Physical activity, 8 2022. Accessed: Aug. 18, 2023.

[38] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023.

[39] Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilinmed-vl: Towards chinese large vision-language model for general healthcare. arXiv preprint arXiv:2310.17956, 2023.

[40] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676, 2019.

[41] Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, John Wilbur, and Zhiyong Lu. Biocpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. arXiv preprint arXiv:2307.00589, 2023.

[42] 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.

[43] A Venigalla, J Frankle, and M Carbin. Biomedlm: a domain-specific large language model for biomedical text. MosaicML. Accessed: Dec, 23(3):2, 2022.

[44] Weihao Gao, Zhuo Deng, Zhiyuan Niu, Fuju Rong, Chucheng Chen, Zheng Gong, Wenze Zhang, Daimin Xiao, Fang Li, Zhenjie Cao, Zhaoyi Ma, Wenbin Wei, and Lan Ma. Ophglm: Training an ophthalmology large language-and-vision assistant based on instructions and dialogue, 2023.

[45] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. Construction of the literature graph in semantic scholar. arXiv preprint arXiv:1805.02262, 2018.

[46] Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, et al. Meddialog: a large-scale medical dialogue dataset. arXiv preprint arXiv:2004.03329, 3, 2020.

[47] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. arXiv preprint arXiv:2311.10537, 2023.

[48] Toyhom. Données de dialogue médical en chinois. https://github.com/Toyhom/Chinese-medical-dialogue-data, 2023. Dépôt GitHub.

[49] 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.
[50] https://www.icliniq.com/
[51] Odma Byambasuren, Yunfei Yang, Zhifang Sui, Damai Dai, Baobao Chang, Sujian Li, and Hongying Zan. Preliminary study on the construction of chinese medical knowledge graph. Journal of Chinese Information Processing, 33(10):1–9, 2019.
[52] Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. arXiv preprint arXiv:2308.03549, 2023.
[53] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454, 2023.