Unlocking the Black Box? A Comprehensive Exploration of Large Language Models in Rehabilitation

Peer-reviewed author version

BONNECHERE, Bruno (2024) Unlocking the Black Box? A Comprehensive Exploration of Large Language Models in Rehabilitation. In: American journal of physical medicine & rehabilitation, 103 (6) , p. 532 -537.

DOI: 10.1097/PHM.0000000000002440
Handle: http://hdl.handle.net/1942/43329

Unlocking the Black Box? A Comprehensive Exploration of Large Language Models in Rehabilitation

Bruno Bonnechère[1,2,3]

[1] REVAL Rehabilitation Research Center, Faculty of Rehabilitation Sciences, Hasselt University, Diepenbeek, Belgium.
[2] Technology-Supported and Data-Driven Rehabilitation, Data Sciences Institute, Hasselt University, Diepenbeek, Belgium.
[3] Department of PXL – Healthcare, PXL University of Applied Sciences and Arts, Hasselt, Belgium

Corresponding author: Prof. dr. Bruno Bonnechère – Bruno.bonnechere@uhasselt.be

Agoralaan – Building A

3590 Diepenbeek

Belgium

Author disclosure

The author declares that he has no relevant or material financial interests that relate to the research described in this paper

Data availability

Not applicable

Funding

No funding was received

Funding

No funding was received

Ethical approval

Not applicable

Abstract

Rehabilitation is a vital component of healthcare, aiming to restore function and improve the well-being of individuals with disabilities or injuries. Nevertheless, the rehabilitation process is often likened to a *'black box'*, with complexities that pose challenges for comprehensive analysis and optimization. The emergence of Large Language Models (LLMs) offers promising solutions to better understand this *'black box'*. LLMs excel at comprehending and generating human-like text, making them valuable in the healthcare sector. In rehabilitation, healthcare professionals must integrate a wide range of data to create effective treatment plans, akin to selecting the best ingredients for the *'black box'*. LLMs enhance data integration, communication, assessment, and prediction.

This paper delves into the ground-breaking use of LLMs as a tool to further understand the rehabilitation process. LLMs address current rehabilitation issues, including data bias, contextual comprehension, and ethical concerns. Collaboration with healthcare experts and rigorous validation is crucial when deploying LLMs. Integrating LLMs into rehabilitation yields insights into this intricate process, enhancing data-driven decision-making, refining clinical practices, and predicting rehabilitation outcomes. Although challenges persist, LLMs represent a significant stride in rehabilitation, underscoring the importance of ethical use and collaboration.

Keywords

Large Language Model; black-box; Rehabilitation; Research; Methods

Introduction

The global demographic landscape, marked by an aging population and a rising incidence of non-communicable diseases and injuries, has led to a growing number of individuals facing disabilities or declines in functionality.[1] This demographic and health transition emphasizes the imperative for health policy planners to prioritize rehabilitation services, which encompass a range of interventions crucial for individuals facing limitations in physical, mental, and social functioning. Rehabilitation proves both effective as an intervention and cost-effective, cutting hospitalization expenses and reducing lengths of stay. The integration of advanced technological and digital solutions, such as online programs and assistive technologies has become prevalent in rehabilitation, benefiting millions globally.[1] Recognizing the complexity of the rehabilitation process, referred to as a *'black box'*,[2] involving patient-specific characteristics, diverse therapeutic approaches, treatment outcomes, and subjective elements, healthcare professionals and researchers have actively sought innovative approaches to unravel these intricacies. However, several constraints within this domain have impeded progress. One notable limitation arises from the lack of validation due to the absence of standardization and precise definitions of interventions in rehabilitation.[3] The personalized nature of treatments, tailored to individual patient needs, results in a dynamic and variable treatment landscape, hindering replicability in randomized controlled trials and establishment of evidence-based practices in rehabilitation. A second limitation stems from the subjective evaluation of outcomes by clinicians, raising concerns about the sensitivity of assessments to subtle patient modifications. Traditional quantitative outcomes may not fully capture the complexities of real-world activities of daily living, creating a gap between research findings and practical applicability. This limitation underscores the pressing need for more nuanced and sensitive outcome measures in rehabilitation research. Furthermore, the translation of research findings to clinical practice faces hurdles, with low external validity attributed to treatment and selection biases. The challenges in effectively bridging the gap between research and clinical implementation underscore the need for innovative solutions that enhance accuracy in assessment, prognosis, patient selection, and decision-making within clinical practice.

Artificial Intelligence (AI) demonstrates considerable promise in addressing these challenges.[4] Despite exploratory initiatives undertaken to study its applicability in evidence-based medicine,[5] there is still a knowledge gap regarding the specific role AI might play in rehabilitation medicine. This paper explores the potential of employing AI Large Language Models (LLMs) to address translational rehabilitation research issues, aiming to provide accurate assessments, prognosis, patient selection, and decision-making, ultimately elevating the quality of care in rehabilitation.

Large Language Models

Definition

LLMs can be defined as a specialized type of AI that has been trained on vast amounts of text to understand existing content and generate original content. LLMs represent advanced AI systems, leveraging extensive datasets sourced from articles, books, and online content to capture intricate word associations through neural network architectures. Employing deep learning methodologies, these models undergo multi-staged iterations involving varying degrees of human input to discern patterns governing word interactions during training. Proficient in natural language processing (NLP), LLMs emulate human-like linguistic abilities, contributing to the evolution of language analysis automation.[6]

The synergy between AI and healthcare holds promises for enhancing healthcare delivery, improving outcomes, and reducing costs. In medical applications, LLMs follow a development process involving electronic health records (EHRs) notes and other medical documentation. After preprocessing to ensure privacy and accuracy, LLMs generate a vocabulary distribution, facilitating predictions of diagnoses, treatments suggestions, and providing decision support.[7] The versatile applications of

LLMs include generating discharge summaries, extracting clinical concepts, responding to medical queries, interpreting electronic health records, and composing medical articles.

State of the art literature

Currently, LLMs are not positioned to replace doctors or therapists due to their imperfect competence in specialized examinations. However, promising results suggest that existing technology has the potential to influence clinical practice, with further advancements anticipated to accelerate and broaden the applications of NLP AI in Physical Medicine and Rehabilitation (PM&R).[8]

LLM trained for medical tasks achieved comparable performance to traditional approaches in *predicting* re-admission, in-hospital mortality, length of stay, and comorbidity index.[7] While these models have value in healthcare for non-diagnostic tasks like NLP of medical literature, documentation assistance, and providing general information, it is crucial to recognize their limitations and rely on qualified healthcare professionals for medical diagnosis.

In *predicting* seizure recurrence, LLMs surpassed models trained with structured data, demonstrating their potential to enhance predictive capabilities in specific medical domains. A case study using CHATGPT-4 as a clinical tool in rehabilitation medicine showed its effectiveness in formulating rehabilitation prescriptions highlighting potential use in clinical practice and education.[5]

Furthermore, an exploration of LLMs for *therapy recommendations* in ophthalmology, orthopaedics, and dermatology revealed proficiency but raised concerns about content quality and potential harm.[9]

Lastly, a study aiming to evaluate LLMS for generating patient message responses in an EHR portal. The dataset included 499,794 pairs, with input from four primary care physicians.[10] Generated responses received positive evaluations for responsiveness, empathy, and accuracy, with a neutral rating for usefulness, showing promise for enhancing *communication* between patients and healthcare providers.

Potential added value

The potential of LLMs in addressing the aforementioned limitations is significant and wide-ranging. LLMs excel in evaluating large datasets, incorporating existing knowledge, and generating tailored content (see Figure 1).[11] Over the last years, the use of technology-supported rehabilitation has gained popularity. Most of the devices enable continuous motion monitoring during rehabilitation exercises, referred to as biomarkers, encompassing various medical indicators associated with biological processes, pathogenic responses, and interventions. Rehabilomics offers a novel framework for discussing biomarkers in PM&R,[12] involving systematic data collection of rehabilitation-related phenotypes and interdisciplinary analysis of biomarkers to gain insights into the biology, function, prognosis, complications, treatments, adaptation, and recovery of individuals with disabilities.

By using LLMs in *scientific research*, researchers have the potential to analyse the large amount of data generated using data mining, can conduct sophisticated power evaluations, hence improving the efficacy of study designs and ensuring the reliability and strength of the obtained results.[13] Additionally, LLMs have the potential to foster uniformity and reproducibility in research initiatives by facilitating the development of standardized language and descriptions of interventions.[14]

In *clinical settings*, LLMs enhance patient communication via chatbots, making information comprehensible and easily accessible. This fosters patient involvement, collaborative decision-making, and ultimately improved healthcare outcomes.[4] LLMs serve as significant tools for clinical decision support, assisting practitioners in evidence-based decision-making and developing tailored treatment plans. The utilization of rehabilitation technology allows for remote monitoring, which enables LLMs to engage in ongoing evaluation and provide feedback on patients' advancement.[15] This facilitates the implementation of personalized rehabilitation plan, and has the ability to forecast rehabilitation outcomes. The utilization of this technology could boost the precision and effectiveness of rehabilitation assessments, ultimately leading to improved patient outcomes.[16]

Rehabilitation professionals, irrespective of their specific area of expertise, will experience enhanced ease in collaborating with diverse AI applications, such as radiological AI and EHRs administration AI, among others. LLMs will enhance the efficacy of communication and data interchange among diverse AI systems, hence enabling the dissemination of findings and insights from one domain to another.

Current limitations, challenges and pitfalls

While LLMs offer promising solutions for translational rehabilitation research and clinical practice, it is crucial to acknowledge that they come with their own set of limitations and potential pitfalls. Table 1 and Figure 2 outline these challenges, spanning technical obstacles, as data diversity, interoperability, data privacy, and security, and ethical considerations.[17]

Moreover, LLMs may lack the specificity and context sensitivity needed for intricate rehabilitation assessments and interventions. The intricacies of rehabilitation require a multifaceted understanding of individual patient needs, nuanced contextual factors, and real-time adjustments based on patient progress. Despite LLMs' proficiency in certain predictive medical tasks, they may not possess the depth of understanding and adaptability necessary for personalized rehabilitation planning. Hence, while recognizing the utility of LLMs in broader healthcare applications, it's crucial to emphasize their supplementary role in the rehabilitation domain. Qualified healthcare professionals, with their expertise and personalized insights, remain indispensable for conducting comprehensive rehabilitation assessments and devising tailored intervention strategies that consider the unique nuances of each patient's journey toward recovery.

The interoperability issue is underscored by significant variations in content quality and safety observed across different LLMs and specialties.[9] The choice of training data for the LLM is a critical consideration. Training on local EMRs captures specific nuances of local practices and terminology, providing a tailored model that aligns closely with the characteristics of a particular healthcare setting. On the other hand, training on a highly varied set of data aims to enhance interoperability, enabling the model to adapt to diverse healthcare contexts. However, this approach raises questions about the need for standardizing the documentation lexicon. Striking a balance between specificity and interoperability is crucial. While standardization can facilitate model understanding across different systems, it may not fully capture the richness of local practices. It is a complex trade-off, and the feasibility of handling idiosyncratic variations depends on the model's adaptability and the level of standardization required for effective communication and interoperability across healthcare settings.

One of the most important challenge in rehabilitation arises from the diverse and often uninformative terminology used to describe treatments across different disciplines. While objective performance data, demographic information, and certain outcome measures may provide more interpretable inputs to LLMs, the documentation of treatments in Electronic Medical Records (EHR) poses a significant hurdle. Current EMR documentation tends to emphasize the duration and intensity of treatment rather than its underlying mechanisms. This documentation approach makes it challenging for LLMs to capture the nuanced details of therapeutic interventions delivered by human therapists. Moreover, the absence of documentation on certain issues in patient descriptions may not necessarily indicate their absence but could signify a lack of focus due to resource constraints and the necessity to prioritize specific aspects of treatment. Addressing these challenges requires a nuanced understanding of the intricacies within the language used to document rehabilitation processes and a careful consideration of the limitations in the current documentation practices.[7,13,14]

The fundamental importance lies in guaranteeing the responsible and ethical utilization of LLMs within the healthcare domain. In order to fully leverage the capabilities of LLMs, it is imperative to address and overcome the various technological challenges that arise. As the field is still evolving and maturing, the fundamental principles of data management within rehabilitation also need to be addressed, captured in the FAIR framework, which stands for Findability, Accessibility, Interoperability, and Reuse of digital datasets (FAIR).[18] This framework requires the creation of unique and deidentified metadata for easy discovery, open or federated access points for accessibility, comprehensive data sharing for interoperability, and data with accurate attributes under clear usage agreements for reusability. Incorporating LLMs into rehabilitation, alongside the FAIR framework and rehabilomics, provides valuable insights into the previously complex rehabilitation process.

Moreover, the willingness of the rehabilitation community to accept and adjust to novel technologies such as LLMs is crucial for their effective integration. The effective and responsible integration of LLMs in the field of rehabilitation necessitates the crucial involvement of technology specialists, physicians, and researchers working collaboratively to overcome these limits. Therefore, in order to optimize the advantages of LLMs while addressing their limitations and drawbacks, it is imperative to employ them as instruments in hybrid way with the expertise of clinicians,[19] thoroughly assess their results, and maintain ethical and regulatory compliance throughout their utilization in translational rehabilitation research.

Discussion

The incorporation of LLMs into the field of rehabilitation presents significant opportunities for elucidating the complex nature of this process. The use of LLMs, for the examination and comprehension of rehabilitation data, holds the potential to augment knowledge, improve clinical decision-making, and optimize rehabilitation outcomes. Despite the existence of ongoing limitations and challenge such as data diversity, interoperability, data privacy and security, the adoption to medical knowledge, innovation readiness and user acceptance, it is imperative not to disregard the significant potential advantages associated with this ground-breaking methodology. The progression from unimodal to multimodal AI represents an essential and imperative advancement in order to effectively exploit the complete capabilities of AI in the healthcare. A ten-point strategy to evaluate and ease the implementation of LLM in rehabilitation is proposed in Table 2, summarizing current recommendations.[7,9,14,20]

While LLMs have the potential to be valuable tools in rehabilitation research and practice, it's important to note that they are not a replacement for traditional methods and the expertise of healthcare professionals. They should be used as supplements to human decision-making and research processes. Additionally, collaboration between researchers, clinicians, and technology experts is essential to harness the full potential of language models in addressing the limitations in rehabilitation research.

References

1. Cieza A, Causey K, Kamenov K, Hanson SW, Chatterji S, Vos T. Global estimates of the need for rehabilitation based on the Global Burden of Disease study 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2021;396(10267):2006-2017. doi:10.1016/S0140-6736(20)32340-0

2. Whyte J, Hart T. It's more than a black box; it's a Russian doll: defining rehabilitation treatments. *Am J Phys Med Rehabil*. 2003;82(8):639-652. doi:10.1097/01.PHM.0000078200.61840.2D

3. Van Stan JH, Dijkers MP, Whyte J, et al. The Rehabilitation Treatment Specification System: Implications for Improvements in Research Design, Reporting, Replication, and Synthesis. *Arch Phys Med Rehabil*. 2019;100(1):146-155. doi:10.1016/j.apmr.2018.09.112

4. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. Drazen JM, Kohane IS, Leong TY, eds. *N Engl J Med*. 2023;388(13):1233-1239. doi:10.1056/NEJMsr2214184

5. Zhang L, Tashiro S, Mukaino M, Yamada S. Use of artificial intelligence large language models as a clinical tool in rehabilitation medicine: a comparative test case. *J Rehabil Med*. 2023;55:jrm13373. doi:10.2340/jrm.v55.13373

6. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl*. 2023;82(3):3713-3744. doi:10.1007/s11042-022-13428-4

7. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194. doi:10.1038/s41746-022-00742-2

8. Thirunavukarasu AJ, Hassan R, Mahmood S, et al. Trialling a Large Language Model (ChatGPT) in General Practice With the Applied Knowledge Test: Observational Study Demonstrating Opportunities and Limitations in Primary Care. *JMIR Med Educ*. 2023;9:e46599. doi:10.2196/46599

9. Wilhelm TI, Roos J, Kaczmarczyk R. Large Language Models for Therapy Recommendations Across 3 Clinical Specialties: Comparative Study. *J Med Internet Res*. 2023;25:e49324. doi:10.2196/49324

10. Liu S, McCoy AB, Wright AP, et al. *Leveraging Large Language Models for Generating Responses to Patient Messages*. Health Informatics; 2023. doi:10.1101/2023.07.14.23292669

11. Seth I, Lim B, Xie Y, et al. Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study. *Aesthet Surg J Open Forum*. 2023;5:ojad084. doi:10.1093/asjof/ojad084

12. Wagner AK. A Rehabilomics framework for personalized and translational rehabilitation research and care for individuals with disabilities: Perspectives and considerations for spinal cord injury. *J Spinal Cord Med*. 2014;37(5):493-502. doi:10.1179/2045772314Y.0000000248

13. Fernandes MB, Valizadeh N, Alabsi HS, et al. Classification of neurologic outcomes from medical notes using natural language processing. *Expert Syst Appl*. 2023;214:119171. doi:10.1016/j.eswa.2022.119171

14. Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. *Nat Hum Behav*. 2023;7(9):1526-1541. doi:10.1038/s41562-023-01659-w

15. Adans-Dester C, Hankov N, O'Brien A, et al. Enabling precision rehabilitation interventions using wearable sensors and machine learning to track motor recovery. *NPJ Digit Med*. 2020;3(1):121. doi:10.1038/s41746-020-00328-w

16. Lin DJ, Stein J. Stepping Closer to Precision Rehabilitation. *JAMA Neurol*. 2023;80(4):339-341. doi:10.1001/jamaneurol.2023.0044

17. Watters C, Lemanski MK. Universal skepticism of ChatGPT: a review of early literature on chat generative pre-trained transformer. *Front Big Data*. 2023;6:1224976. doi:10.3389/fdata.2023.1224976

18. Sinaci AA, Núñez-Benjumea FJ, Gencturk M, et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. *Methods Inf Med*. 2020;59(S 01):e21-e32. doi:10.1055/s-0040-1713684

19. Fu S, Thorsteinsdottir B, Zhang X, et al. A hybrid model to identify fall occurrence from electronic health records. *Int J Med Inform*. 2022;162:104736. doi:10.1016/j.ijmedinf.2022.104736

20. Perlis RH, Fihn SD. Evaluating the Application of Large Language Models in Clinical Research Contexts. *JAMA Netw Open*. 2023;6(10):e2335924. doi:10.1001/jamanetworkopen.2023.35924

Figures caption

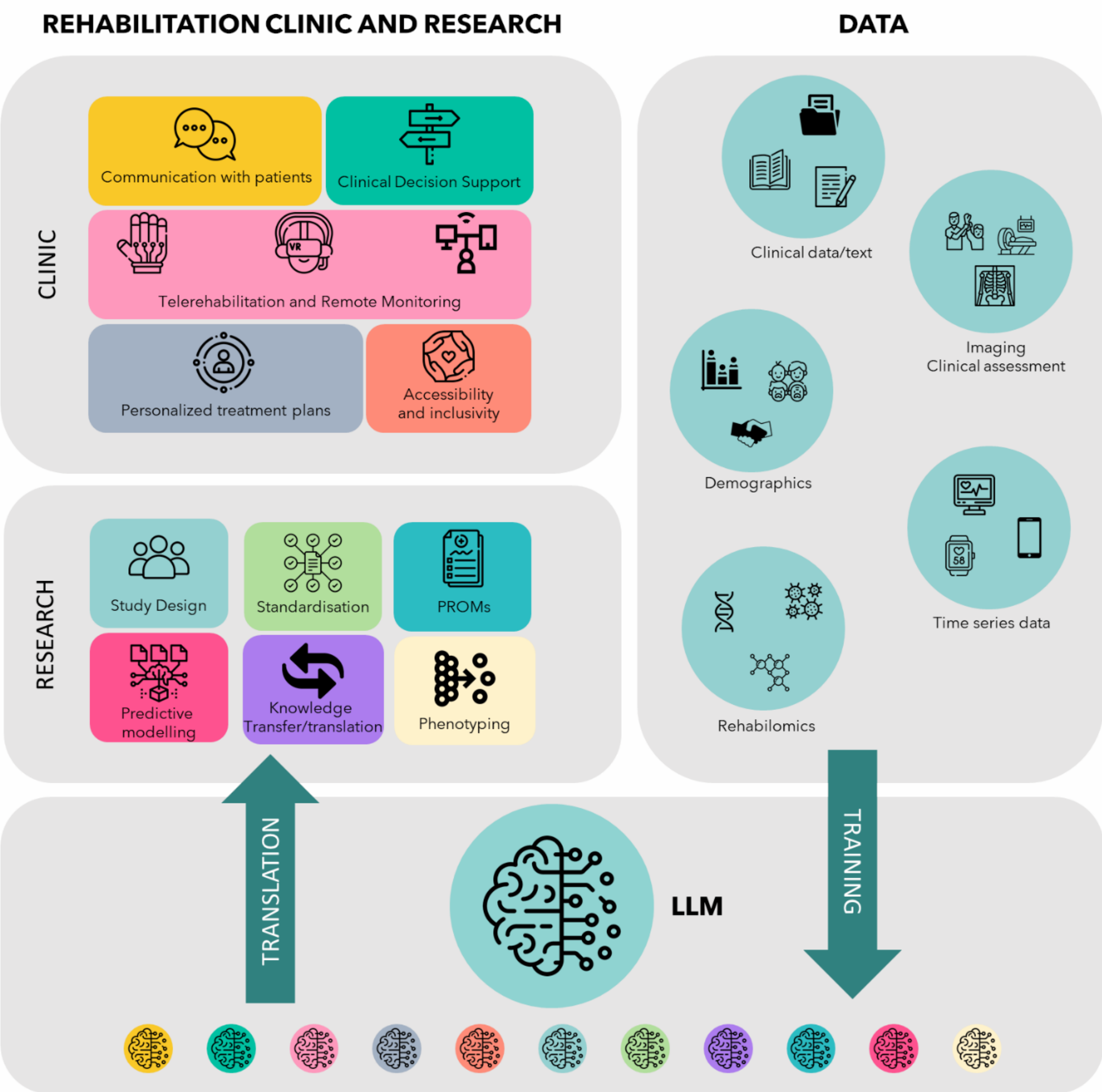

Figure 1: LLMs in Rehabilitation Research and Clinical Practice

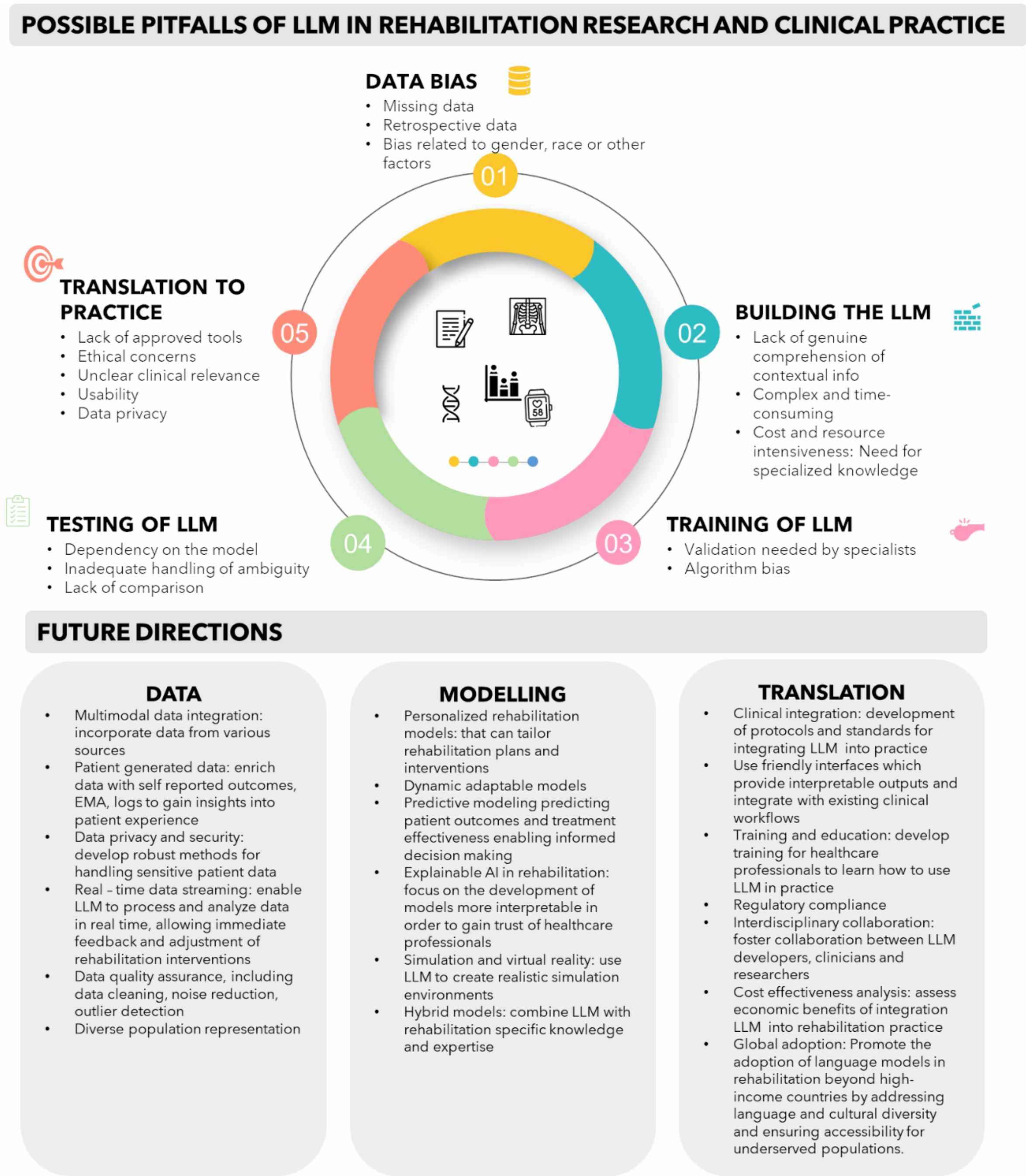

Figure 2: Pitfalls and Future Directions of LLM in Rehabilitation Research and Clinical Practice

Table 1: Ten-Key challenges in using LLMs in rehabilitation research and clinical practice

Challenges	Definition
Data bias	LLMs are trained on large and diverse datasets, but these datasets can still contain biases present in the text from which they learn. If the training data includes biases related to gender, race, or other factors, LLM-generated content may inadvertently perpetuate these biases. By definition the models are trained using retrospective data, we therefore have little control on the quality of these
Lack of contextual understanding	LLMs employ pattern recognition techniques to generate text by analysing data. However, it is important to note that LLMs lack genuine comprehension of contextual information. The aforementioned phenomenon has the potential to give rise to seemingly credible yet inaccurate or deceptive information, particularly within intricate scientific domains such as rehabilitation
Ethical concerns	The utilization of LLMs gives rise to ethical considerations pertaining to the conscientious application of technology. The advent of automated content generation and information manipulation holds the potential to give rise to issues such as disinformation, plagiarism, and various other unethical actions
Over reliance on automation	Excessive dependence on LLMs can potentially result in a decline in critical thinking abilities and human knowledge. There is a possible inclination among researchers to completely replace human-generated content with LLM-generated content, which might potentially compromise the quality of research and patient care
Data privacy	The use of LLMs frequently entails the handling and analysis of confidential patient data and health-related information. The

	maintenance of data privacy and adherence to pertinent legislation are of utmost importance in order to mitigate breaches and legal complications
Validation	The content produced by LLMs, particularly in healthcare sector, necessitates thorough validation by specialists in the respective domain. Failure to adhere to this practice may lead to the dissemination of erroneous or untrustworthy data
Cost and resource intensiveness	The process of creating, refining, and sustaining Language Models (LLMs) might require significant resources in terms of time, specialized knowledge, and computational capabilities, making it impractical for certain research teams
Dependency on technology	An overreliance on LLMs can hinder researchers' critical thinking skills and creativity, potentially leading to a lack of innovation in problem-solving
Algorithmic bias	The utilization of LLMs has the potential to mistakenly produce biased material that mirrors the inherent biases within the training data. This can pose challenges in situations that necessitate impartiality and equity
Inadequate handling of ambiguity	LLMs struggle with handling ambiguity and uncertainty. The tendency observed is that individuals often offer comments that convey a sense of confidence, even in situations where the input may possess ambiguities or allow for many correct interpretations. The aforementioned constraint can provide significant challenges, particularly when tackling intricate and multifaceted inquiries in the field of translational rehabilitation research, where accurate responses are of utmost importance

Table 2: Ten-Point Strategy for Evaluating LLM Applications in Rehabilitation Research

Strategy	Description
Clinical Significance	Ensure that the utilization of LLM in rehabilitation research effectively addresses pertinent therapeutic aspects. Reports should cover a wide range of experimental conditions that are relevant to clinical settings, while avoiding a limited scope.
Detailed experimental conditions	Present a comprehensive account of the experimental settings, encompassing adjustable factors such as metadata and data used to train the model, to facilitate replication by proficient users. Indicate the specific version of the LLM, acknowledging that they can undergo changes and improvements over time (see Point 4).
Stability through replication	Perform numerous repetitions to evaluate the consistency of LLM responses. Repeating requests is essential in order to accurately estimate impacts and comprehend the variability, considering the probabilistic nature of text generation.
Comparison of LLM versions	Highlight the assessment of the latest iterations of LLM models, as comparisons with previous iterations may have limited relevance to individuals without specialized knowledge. Emphasize enhancements and progress made in the most recent versions.
Characteristics of incorrect responses	Go beyond fundamental accuracy measures to delineate attributes of erroneous or defective answers. Offer analysis on the circumstances under which LLMs can produce erroneous outcomes, substantiated with illustrations and practical examples.
Assessment of bias and fairness	Systematically assess bias and fairness in LLM applications. Acknowledge the susceptibility to biases and report efforts made to minimize them. Consider modifications to prompts to address individual characteristics.
Confidentiality	Ensure the protection of sensitive information when utilizing LLMs for clinical data. Provide a detailed description of the steps

protection	implemented, such as hosting LLMs within institutional firewalls, to ensure that unauthorized individuals cannot access the data and to uphold the confidentiality of patient information.
Comparison to expert reference	Evaluate the LLM outcomes by contrasting them with recognized expert benchmarks. Promote the incorporation of these reference standards in the reporting, particularly those involving physicians or human subjects, which necessitate ethical evaluation.
Reproducibility	Enhance the ability to replicate experiments utilizing LLMs by offering comprehensive instructions on promoting reproducibility. Provide details regarding parameters, datasets, and settings to enable autonomous verification of outcomes (see Point 2).
User accessibility	Examine the usability of LLM apps for end-users, specifically healthcare professionals. Ensure that the results are easily understandable and useful in making healthcare decisions, promoting user acceptability and participation.