Volume 4, Issue 6, 2026

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

LARGE LANGUAGE MODELS AND GENERATIVE AI IN PHYSIOTHERAPY CLINICAL WORKFLOWS

Dr Iqra Taj (PT)[1], Dr Mamoona Waseem (PT)[2], Dr Faiza Waheed (PT)[3], Dr Alwaaz Zeeshan (PT)[4], Dr Tooba Shahbaz (PT) [5], Dr. Inayat Ullah (PT)[*6]

[1]*Student, MS Cardiopulmonary Physical Therapy, FUCP, Foundation University, Islamabad*
[2]*Clinical Physical Therapist, PhysioPlus Clinic, Peshawar*
[3]*Demonstrator, Foundation University, Islamabad*
[4]*Physiotherapist, PAF Hospital, Islamabad*
[5]*Physiotherapist, Fauji Foundation Hospital, Rawalpindi*
[*6]*Assistant Professor, Sarhad University of Science & Information Technology Peshawar*

[1]iqisraran@gmail.com, [2]Mamoona855@gmail.com, [3]faiza.waheed@fui.edu.pk, [4]alwaaz.zeeshanf20@gmail.com, [5]tshahbaz59@gmail.com, [*6]inayatullah.siahs@suit.edu.pk

DOI:https://doi.org/10.5281/zenodo.20771306

Keywords
large language models; generative artificial intelligence; ChatGPT; physiotherapy; clinical documentation; patient education; scoping review

Article History
Received: 22 April 2026
Accepted: 04 June 2026
Published: 20 June 2026

Copyright @Author

Corresponding Author: *
Dr. Inayat Ullah

Abstract

Background: *For documentation, patient education, clinical reasoning support, and professional education, large language models (LLMs) such as ChatGPT, Gemini, and Claude are widely suggested. Physiotherapy-specific evidence is becoming more prevalent yet dispersed. The aim is to map the nature, extent, and gaps of the literature on LLM/generative-AI applications in physiotherapy clinical and clinical-adjacent workflows.*

Methodology: *Indexed literature (January 2022-June 2025, with practical yield from late 2022) was searched and screened; a scoping review was carried out in accordance with JBI and reported in accordance with PRISMA-ScR. Seven sources were charted and mapped by application domain.*

Results: *There are three main application clusters: (1) text generation, patient education and documentation, where models can create, simplify, and translate materials; (2) professional education, which includes clinical cases generated by LLM for students; and (3) clinical-reasoning benchmarking, which compares models to clinicians in particular rehabilitation domains. Although task-dependent and prone to errors, accuracy is promising, and there is a dearth of deployed clinical evaluations as opposed to instructional or benchmark use. Non-English ethical, governance, privacy, and equity issues are frequently brought up yet remain unresolved.*

Conclusion: *Supporting documentation and creating patient education materials are the most evidence-based applications; clinical reliance necessitates local validation and governance. The main deficiency is in the deployment of multilingual, physiotherapy-specific evidence.*

INTRODUCTION

In addition to drafting documents, summarising, responding to enquiries, and simulating conversation, generative language models offer a natural-language interface to massive text corpora. Early medical syntheses identified problems related to accuracy, bias, privacy, and accountability while suggesting roles in patient

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

education, recordkeeping, office administration, and decision support (1). A thorough analysis of ChatGPT as a clinical support tool categorised its use into three areas: therapeutic/patient services, clinical decision support, and data management (including documentation workflows). It was observed that dialogue-based frameworks had special potential in physiotherapy (2).

The body of research in physiotherapy is growing unevenly. In addition to the large number of patient education demands for home programmes and self-management, the profession faces unique documentation requirements, such as daily notes, exercise programme instructions, and progress summaries. LLMs may provide instantaneous efficiency advantages in some text-generation jobs. However, clinical reasoning in physiotherapy is extremely contextual, depending on patient conversation, movement observation, and hands-on assessment; domains in which LLMs do not receive direct sensory information. This generates a particular tension. The most commonplace (drafting text) may be the most instantly helpful, while the most advertised (reasoning support) may be the least reliable. From creating educational contexts for students (3) to comparing LLM performance on vestibular knowledge tests, early discipline-specific studies have started examining both extremes of this continuum (4). Nevertheless, no previous synthesis has delineated the entire spectrum of LLM applications across clinical and clinical-adjacent activities in physiotherapy.

Review Questions

The following questions were attempted to answer in this scoping review:

1. In which physiotherapy workflows have LLMs been applied or evaluated?
2. What models, designs, tasks, and sub-specialities are reported?
3. How are performance and accuracy assessed, against what comparators?
4. What safety/ethical/governance/equity issues and mitigations are reported?
5. Where are the principal gaps, especially for deployed clinical and non-English use?

Methods

On May 1, 2026, a structured search of the published literature (PubMed/MEDLINE and related indexed sources) was carried out for this review. Title/abstract and full-text screening were done in accordance with the eligibility criteria listed below. The evidence tables and flow counts show the results of the search.

Population: Physiotherapists, physiotherapy students/educators, and physiotherapy patients/settings; mixed-profession studies are included where physiotherapy data are separable and flagged as indirect.

Concept: Application or evaluation of LLMs/generative AI (ChatGPT, GPT-4, Gemini, Claude, etc.) in clinical or clinical-adjacent physiotherapy workflows.

Context: Any clinical, academic, or community setting in any country. Records January 2022–June 2025.

Results

Search and selection

Seven sources were charted after screening (Table 1): three physiotherapy-specific evaluations (musculoskeletal clinical reasoning and vestibular/comparative benchmarking), one discipline-specific educational study, two general-medicine syntheses with physiotherapy relevance, and one narrative review in paediatric physiotherapy. The PRISMA-ScR flow diagram is provided in Fig 1.

Volume 4, Issue 6, 2026

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

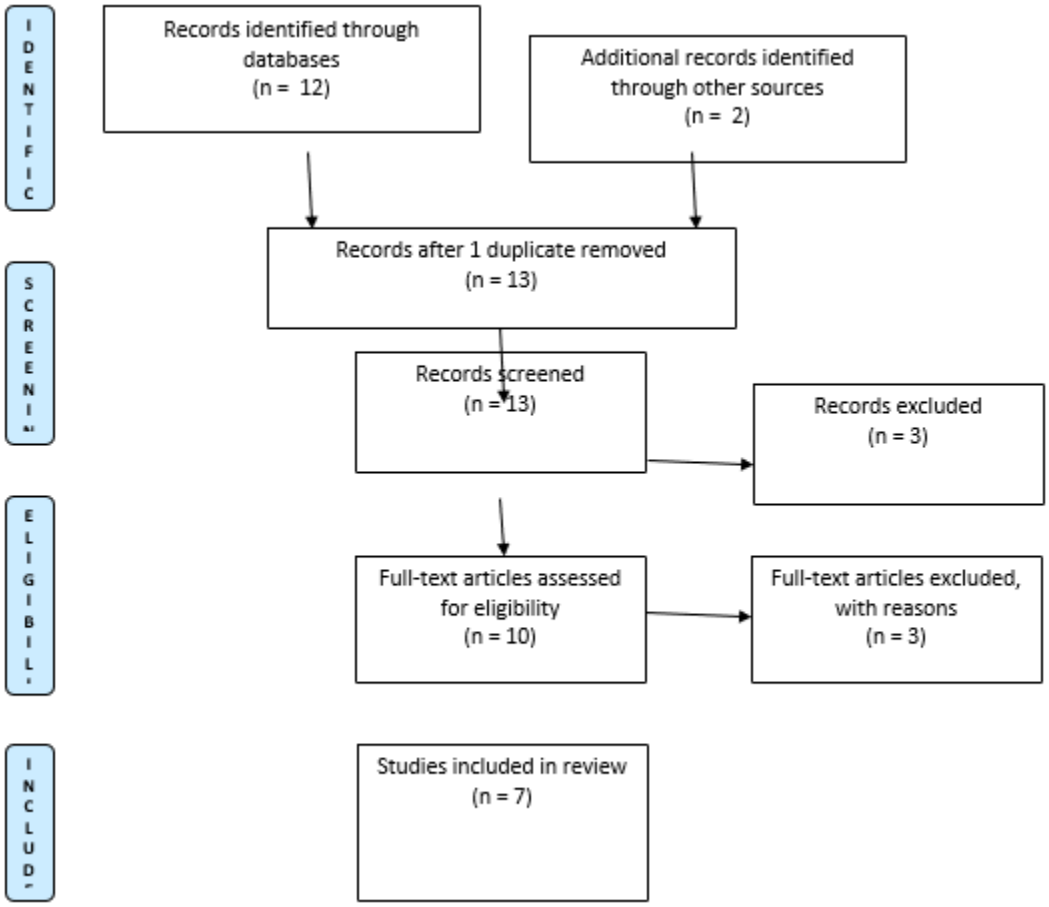


Figure 1: PRISMA flowchart of the selection of studies

Documentation and patient education (Q1, Q3)

Text-generation jobs are the most evidence-based applications. A scoping review of LLMs in patient education across medicine revealed their use for creating, streamlining, translating, and improving materials; for instance, lowering the reading level and using passive voice to increase health literacy accessibility; functions that directly translate to physiotherapy home programmes and self-management content (2). General clinical-support syntheses also list record integration and documentation workflow optimisation as top data-management use cases, with physiotherapy specifically mentioned as a promising dialogue-based setting (1, 5).

Professional education and clinical reasoning (Q1, Q2, Q3)

The focus of discipline-specific evidence is on benchmarking and education. An educational project that generated longitudinal respiratory cases for physiotherapy students using an LLM (GPT-4) inside a story-centred curriculum reported positive learning impact and engagement (3). While warning about openness and factual reliability, a narrative review in paediatric physical therapy delineated responsibilities across clinical rationale, customised therapeutic recommendations, and carer education (4). A comparison research in vestibular rehabilitation assessed ChatGPT and Gemini against physiotherapists and students on a structured knowledge test (6), while evaluations of clinical-reasoning competence in musculoskeletal

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

treatment benchmarked ChatGPT against the anticipated standard (7). Accuracy was promising, but task-dependent and inconsistent.

Safety, ethics, governance, and equity (Q4, Q5)

Factual mistakes and "hallucinations", opacity, algorithmic bias, data privacy, medico-legal accountability, and erosion of the relational/humanistic parts of care were among the governance issues raised by all of the included sources; nevertheless, mitigations were suggested rather than assessed (1, 4, 5). The majority of the research focuses on instructional or benchmark settings rather than practical clinical use with patient outcomes, and it seldom looks at non-English/lower-resource performance, which directly limits South Asian practice and creates a priority gap.

Table 1. Charted sources and extracted findings

Study (first author, year)	Type	Focus	Key extracted findings
Kantor, 2023 (1)	Primer/review	LLMs in medicine	Roles: patient education, documentation, office management, education, decision support; accuracy/bias/privacy caveats.
Aydin, 2024 (2)	Scoping review	LLMs in patient education	Generate/simplify/translate/enhance materials; reduce reading level and passive voice.
Okuno, 2025 (3)	Educational study	PT students (respiratory)	GPT-4-generated longitudinal cases in a story-centred curriculum; positive learning impact.
Ziab, 2025 (4)	Narrative review	Paediatric physiotherapy	Clinical reasoning, tailored recommendations, caregiver education; transparency/factual/ethical limits.
Liu, 2025 (5)	Comprehensive review (296 pubs)	ChatGPT clinical support	3-tier framework incl. documentation/record integration; physiotherapy a promising dialogue context.
Arbel, 2025 (6)	Comparative study	Vestibular rehabilitation	ChatGPT vs Gemini vs PTs/students on knowledge test; model accuracy context-dependent.
Hao, 2025 (7)	Evaluation study	Musculoskeletal PT education	Benchmarks clinical-reasoning performance; promising but variable accuracy.

Discussion

This review offers the first physiotherapy-specific map of LLM applications. It demonstrates a distinct maturity gradient: clinical reasoning and decision support uses are still at the educational and benchmarking stage with promising but

Frontier in Medical & Health Research
ISSN: (e) 3007-1607 (p) 3007-1593

inconsistent accuracy, while text-generation tasks (documentation support, patient-education drafting, and simplification) are the most evidence-near and lowest-risk (3, 4, 6, 7). Recent empirical evidence clearly refutes this. Only 33% of responses to a 2025 study assessing ChatGPT-4 on patellar tendinopathy were clinically correct, and the texts needed a college reading level. This demonstrates a substantial risk of clinical misinformation and poor health literacy compliance if expert verification is circumvented, which runs counter to the safety premise (8).

The practical implication for doctors is to treat reasoning and decision outputs as unvalidated and to use LLMs first to create and streamline patient materials and documentation, always with clinical verification. According to a 2024 study, although LLMs offer sufficient descriptive planning, more than 80% of the clinical references they produce to support their conclusions are hallucinations (9). Additionally, a 2026 benchmarking study showed that LLMs perform worse than simple factual recollection when required to produce step-by-step diagnostic reasoning, undermining their usage in complex triage (10).

A more thorough analysis reveals a number of methodological limitations that limit results. First, the majority of the seven plotted sources are either small-scale, single-institution educational evaluations or general-medicine syntheses with tangential physiotherapy relevance. The two benchmarking studies (6, 7) compare LLM outputs to gold-standard responses or clinician performance on organised knowledge tests. However, real-world clinical reasoning, which contains ambiguity, contradicting inputs, and iterative hypothesis revision, is not well approximated by structured assessments. There are no listed studies that assess the use of LLM in actual clinical workflows using patient outcomes as a goal.

This is a field-wide immaturity rather than a failure of individual experiments; what is available are accuracy and feasibility prototypes rather than efficacy trials. Second, Okuno et al.'s educational study (3) indicates favourable learning impact; nevertheless, the measurements used are self-report and engagement rather than objective competency transfer to clinical settings. Third, while raising important ethical concerns, the narrative review in paediatric physical therapy (4) does not test any intervention, leaving mitigations up to interpretation. There is also a key vacuum for South Asian, African, and other multilingual clinical contexts: the majority of research uses English-language prompts and English-language benchmarks without systematically evaluating non-English performance. When considered collectively, the research only lends credibility to the following weak claims: LLMs can provide credible physiotherapy text in controlled settings, but it is still unclear whether or not to use them and under what administration.

Apart from connecting the evident necessity of deployed clinical evaluations with patient outcomes, three particular gaps are worth highlighting. First, no research has looked at the integration of LLM with electronic health records or documentation systems in physiotherapy, which is a key suggested use case. Second, although they are mentioned in all of the included sources, accountability frameworks (who is responsible when LLM-generated exercise guidance causes harm?) are completely missing from the physiotherapy literature. Third, performance across languages, dialects, health literacy levels, and cultural settings of exercise and rehabilitation is not measured, and the equity dimension is virtually completely ignored. This is a primary problem rather than a secondary one for a worldwide profession such as physiotherapy.

A domain-mapped synthesis and PRISMA-ScR alignment are a few strengths; a fast-moving field and the inclusion of indirect, mixed-profession evidence, given the nascent physiotherapy-specific base, are limitations of this review.

Conclusion

Physiotherapy is adopting generative AI more quickly than the body of evidence supporting it. According to the performed map, clinical reasoning is still in the experimental stage; documentation and patient education drafting are the safest present uses, and deployed, multilingual, governed assessment are the top priorities for research and cautious local acceptance.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

References

Kantor J. ChatGPT, large language models, and artificial intelligence in medicine and health care: A primer for clinicians and researchers. JAAD International [Internet]. 2023 Jul 29;13:168–9. Available from: https://doi.org/10.1016/j.jdin.2023.07.011

Aydin S, Karabacak M, Vlachos V, Margetis K. Large language models in patient education: a scoping review of applications in medicine. Frontiers in Medicine [Internet]. 2024 Oct 29;11:1477898. Available from: https://pubmed.ncbi.nlm.nih.gov/39534227/

Okuno S, Kawamitsu K, Yamaguchi T. Implementation and Educational impact of a Story-Centered Curriculum using a large language model: a class on Internal Disorders for physiotherapy students. Cureus [Internet]. 2025 Oct 12;17(10):e94390. Available from: https://doi.org/10.7759/cureus.94390

Ziab HM, Mazbouh R, Siblini F, Nashwan AJ. Transforming Pediatric Physiotherapy: The Role of ChatGPT in Therapy, Limitations, and Ethical Considerations. Cureus [Internet]. 2025 Sep 3;17(9):e91571. Available from: https://pubmed.ncbi.nlm.nih.gov/41049913/

Liu Y, Zhang Y. ChatGPT as a clinical support tool: A comprehensive review of applications, assessment, and implementation challenges. Physiotherapy Practice and Research [Internet]. 2025 Sep 19;47(1):160–72. Available from: https://doi.org/10.1177/22130683251379638

Arbel Y, Gimmon Y, Shmueli L. Evaluating the potential of large language models for vestibular rehabilitation education: a comparison of ChatGPT, Google Gemini, and clinicians. Physical Therapy [Internet]. 2025 Feb 10;105(4). Available from: https://doi.org/10.1093/ptj/pzaf010

Hao J, Yao Z, Siu K. Artificial Intelligence in Physical Therapy Education: Evaluating Clinical Reasoning Performance in Musculoskeletal Care using ChatGPT. Musculoskeletal Care [Internet]. 2025 Jul 31;23(3):e70177. Available from: https://doi.org/10.1002/msc.70177

Deng J, Li L, Oosterhof JJ, Malliaras P, Silbernagel KG, Breda SJ, et al. ChatGPT is a comprehensive education tool for patients with patellar tendinopathy, but it currently lacks accuracy and readability. Musculoskeletal Science and Practice [Internet]. 2025 Jan 31;76:103275. Available from: https://pubmed.ncbi.nlm.nih.gov/39899928/

Sawamura S, Bito T, Ando T, Masuda K, Kameyama S, Ishida H. Evaluation of the accuracy of ChatGPT's responses to and references for clinical questions in physical therapy. Journal of Physical Therapy Science [Internet]. 2024 Jan 1;36(5):234–9. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC11060764/