Original Article

ANATOLIAN CURRENT MEDICAL

DOI: 10.38053/acmj.1746227

Anatolian Curr Med J. 2025;7(5):695-702

Fabricated or accurate? Ethical concerns and citation hallucination in AI-generated scientific writing on musculoskeletal topics

Ertuğrul Safran[1], Adem Çalı[2]

[1]*Department of Physiotherapy and Rehabilitation, Faculty of Health Sciences, Bezmiâlem Vakıf University, İstanbul, Turkiye*
[2]*Department of Occupational Therapy, Faculty of Health Sciences, İstanbul Topkapı University, İstanbul, Turkiye*

Cite this article as: Safran E, Çalı A. Fabricated or accurate? Ethical concerns and citation hallucination in AI-generated scientific writing on musculoskeletal topics. *Anatolian Curr Med J.* 2025;7(5):695-702.

Received: 19.07.2025 ◆ **Accepted**: 10.09.2025 ◆ **Published**: 15.09.2025

ABSTRACT

Aims: Large language models (LLMs) such as ChatGPT are increasingly used in academic and clinical writing. While these tools can generate coherent and domain-specific text, concerns persist regarding the accuracy of their automatically generated references. In musculoskeletal rehabilitation—a field heavily reliant on current evidence—the reliability of citations is especially critical. Yet, systematic evaluations of citation accuracy in AI-generated scientific content are lacking. To evaluate the reference accuracy of scientific texts generated by ChatGPT (GPT-4) in response to musculoskeletal rehabilitation prompts, and to determine whether reference accuracy improves following structured post-generation verification.

Methods: ChatGPT was prompted to generate four scientific paragraphs on musculoskeletal rehabilitation topics (manual therapy, ACL reconstruction, low back pain, and rotator cuff repair), each including 10 references with DOIs. A total of 40 references were analyzed using a 3-point scoring system (0=fabricated, 1=partially correct, 2=fully accurate), which was used to assess citation quality. After initial evaluation, ChatGPT was asked to verify and revise its references. Scores before and after this step were compared descriptively and with Wilcoxon signed-rank tests to assess statistical significance, and effect sizes (r) were calculated to estimate the magnitude of improvement.

Results: Only 7.5% of references were fully accurate in the initial generation, while 42.5% were completely fabricated. The remaining 50% were partially correct. After verification, the proportion of fully accurate references rose to 77.5%. Wilcoxon signed-rank testing confirmed a statistically significant improvement in accuracy across all prompts ($W=561.0$, $p<0.001$, $r=0.60$). The most common errors included invalid DOIs, fabricated article titles, and mismatched metadata.

Conclusion: ChatGPT can generate coherent scientific content, but its initial references are frequently inaccurate or fabricated. Structured post-generation verification significantly improves reference accuracy, as confirmed by statistical testing. These findings suggest that LLMs may be integrated as drafting tools in academic and clinical musculoskeletal contexts, but only when accompanied by strict human-led verification of citations.

Keywords: ChatGPT, artificial intelligence, musculoskeletal rehabilitation, scientific writing, reference accuracy, hallucinated citations

INTRODUCTION

The integration of artificial intelligence (AI) into academic and scientific writing has introduced powerful tools for generating domain-specific content with unprecedented speed. Among these technologies, large language models (LLMs) such as ChatGPT have demonstrated remarkable fluency in producing coherent and technically structured scientific texts across various disciplines, including health sciences and rehabilitation domains.[1-4] However, a critical and underexplored concern lies in the accuracy and reliability of references generated by these models. In musculoskeletal rehabilitation—including physiotherapy, exercise science, and manual therapy—clinicians, researchers, and educators rely heavily on accurate and up-to-date scientific literature to guide evidence-based decision-making. Citation accuracy is therefore not merely a technical detail but a cornerstone of trustworthy clinical, educational, and research practice. When references generated by LLMs are inaccurate or fabricated, the consequences extend beyond academic integrity, posing risks for patient safety, clinical outcomes, and professional training.[5-8]

Recent investigations have increasingly drawn attention to the phenomenon of "hallucinated references," wherein LLMs such as ChatGPT generate citations that appear syntactically correct and contextually appropriate, yet do not correspond to any real or retrievable publication.[9-12] These fabricated references may include convincing combinations

Corresponding Author: Ertuğrul Safran, esafran@bezmialem.edu.tr

This work is licensed under a Creative Commons Attribution 4.0 International License.

of real author names, journal titles, and plausible publication years, making them difficult to detect without thorough verification.[13] This issue poses a significant challenge in health sciences, where evidence-based accuracy is foundational, and its consequences become particularly acute in clinical fields. Misinformation introduced by inaccurate or fabricated references may propagate unnoticed through educational resources, systematic reviews, or even AI-supported clinical tools—potentially influencing clinical recommendations and practice standards.[14,15] While the growing use of AI-based writing tools introduces opportunities for increased efficiency, it also carries inherent risks associated with unverified bibliographic output.[16-19] For example, an erroneous reference regarding the effectiveness of manual therapy could lead to inappropriate treatment planning, while a fabricated citation on exercise prescription might expose patients to ineffective or unsafe interventions. These risks highlight the necessity of rigorous verification when integrating AI-generated content into musculoskeletal rehabilitation research and practice.

In contrast to previous anecdotal reports or general evaluations,[20,21] there remains a lack of systematic research quantifying the accuracy and correction potential of hallucinated references in musculoskeletal-focused scientific content. Addressing this gap is essential to ensuring the responsible integration of LLMs into scholarly and clinical communication. Therefore, the aim of this study was to systematically evaluate the accuracy of references generated by ChatGPT in the context of musculoskeletal rehabilitation and to examine whether post-generation prompting could improve citation reliability.

METHODS

Ethics

This study did not involve human or animal participants. Therefore, ethical approval was not required according to national guidelines. All procedures were carried out in accordance with the ethical rules and the principles of the Declaration of Helsinki.

Study Design

This was a structured observational analysis comparing four prompt conditions used to generate scientific paragraphs and associated reference lists using ChatGPT (GPT-4 model). All prompts were designed to cover topics related to musculoskeletal disorders in the context of rehabilitation and therapeutic strategies.

Prompt Conditions

The prompts provided to ChatGPT included a consistent instruction to generate a scientific paragraph (approximately 300–400 words) along with 10 references including DOIs.

The following prompts—each of which addressed a clinical topic within the musculoskeletal rehabilitation field—were used:

Prompt 1: Write a short scientific text (approximately 300–400 words) on effects of manual therapy on pain, including 10 references with DOIs.

Prompt 2: Write a short scientific text (approximately 300–400 words) on ACL reconstruction rehabilitation strategies, including 10 references with DOIs.

Prompt 3: Write a short scientific text (approximately 300–400 words) on low back pain exercise strategies, including 10 references with DOIs.

Prompt 4: Write a short scientific text (approximately 300–400 words) on rotator cuff repair rehabilitation strategies, including 10 references with DOIs.

The four clinical topics were deliberately chosen to reflect both breadth and clinical relevance within musculoskeletal rehabilitation. One topic (manual therapy and pain) represents a broad theme that is central across many rehabilitation domains, while the others were selected to cover frequently encountered conditions at different anatomical levels: lower extremity (ACL reconstruction), upper extremity (rotator cuff repair), and spine (low back pain). This distribution allowed for a representative sampling of musculoskeletal rehabilitation contexts without narrowing the analysis to a single region. The instruction to generate 10 references per prompt was used to standardize outputs and to reflect a realistic academic scenario, as users commonly request short scientific texts accompanied by a concise list of approximately ten references. This limit also ensured that the reference lists were manageable for systematic verification while minimizing variability due to overly extensive outputs.

The prompts were administered sequentially in the fixed order listed above (Prompt 1 through Prompt 4). The sequence was not randomized, as the study design aimed to compare outcomes across consistent and replicable conditions. Each prompt was entered in a separate, context-cleared session to ensure that ChatGPT's outputs were not influenced by prior responses.

Following the generation of each response, ChatGPT was given an additional instruction to review the same list of references it had just generated. The exact wording of this post-verification prompt was: 'Please check the references and DOIs you have just generated. Indicate whether each reference exists, whether the metadata (authors, title, journal, year) is correct, and whether the DOI is functional. Correct any errors if possible.' This prompt did not request a completely new reference list; rather, it asked the model to re-examine and, where possible, correct the citations it had already provided. This procedure aimed to simulate a self-correction process and to assess whether the model could improve reference accuracy upon reevaluation. All outputs were generated using ChatGPT (GPT-4, OpenAI web interface, accessed in June 2025). At the time of data collection, OpenAI did not provide a publicly available version number beyond the model designation (GPT-4); therefore, the model is reported with its platform and date of access for transparency.

Reference Evaluation

Each generated reference was evaluated based on the following criteria:

Existence: Does the reference exist in the scientific literature?

Correct author(s): Are the cited authors correct?

Correct title: Is the article title accurate?

Correct journal information: Are the journal name, volume, issue, and pages accurate?

Valid DOI: Does the DOI link resolve to the correct article?

A scoring system was used:

0=Fabricated reference or completely mismatched information

1=Partially correct (e.g., wrong journal or year, incorrect DOI)

2=Fully accurate (exact match across all fields)

In this study, one researcher was responsible for prompt design and content generation, while another independently evaluated the accuracy of the references and DOIs. Both researchers had over 10 years of professional experience in musculoskeletal rehabilitation and academic writing. When DOIs provided by ChatGPT did not resolve automatically, references were verified manually using Crossref, PubMed, Google Scholar, and publisher websites. This predefined protocol ensured that fabricated or mismatched DOIs could be identified and corrected when possible.

Sample Size and Statistical Analysis

A total of 40 references were analyzed (10 references per prompt×4 prompt conditions). Descriptive statistics were used to compare the accuracy scores between prompt conditions. The frequency of score categories (0, 1, 2) was presented as percentages. Additionally, error types were qualitatively categorized (e.g., fabricated reference, wrong DOI, mismatched journal). To further assess changes in reference accuracy before and after verification, Wilcoxon signed-rank tests were performed, with effect sizes (r) calculated to indicate the magnitude of improvement. Statistical significance was set at $p<0.05$. All analyses were conducted using IBM SPSS Statistics 25.

RESULTS

Prompt 1: Effects of Manual Therapy on Pain

A total of 10 references generated in response to prompt 1 were evaluated. None of the references received a full score of 2. The majority of the references (8 out of 10) were partially accurate, receiving a score of 1, typically due to incorrect or missing DOIs, publication years, or mismatched journal details. Two references (20%) were entirely fabricated or had major mismatches (score=0). A detailed breakdown of each reference and its specific issues is presented in **Table 1**.

After prompting ChatGPT to review and verify the accuracy of the references and DOIs, a substantial improvement was observed. Seven references (70%) received a fully accurate score of 2 upon regeneration and verification. Two references (20%) remained partially correct (score=1), and one reference (10%) was still found to be fabricated or completely mismatched (score=0).

Table 1. Initial accuracy evaluation of references generated by prompt 1 (manual therapy on pain)

Reference	Score	Explanation
1	1	DOI and year are incorrect
2	0	Same title exists but by different authors; DOI is non-functional
3	1	DOI and year are incorrect
4	1	Author and topic are similar, but title and journal are different
5	1	DOI is incorrect; author list partially matches a real article
6	1	DOI and year are incorrect
7	1	DOI non-functional; journal name incorrect
8	1	DOI missing; all other citation details are correct
9	0	Fabricated title and DOI under a real author's name
10	1	Correct title and author; journal name and DOI incorrect

DOI: Digital object identifier. DOI validation was conducted manually using Crossref and PubMed, with supplementary checks through Google Scholar when necessary (accessed June 2025).

Prompt 2: ACL Reconstruction Rehabilitation Strategies

Ten references generated in response to prompt 2 were evaluated. None received a fully accurate score of 2 on the initial evaluation. Five references (50%) were partially accurate (score=1), typically due to incorrect DOIs or mismatched journal details, although the articles themselves existed and were relevant to the topic. The remaining five references (50%) were assigned a score of 0, either because they included fabricated article titles, non-functional DOIs, or mismatched information under real author names.

Fabrication was a recurring issue: multiple references contained fabricated article names attributed to real authors who had published similar work in the same year. These cases suggest a pattern in which the AI model tends to blend existing authorship with invented or misattributed titles. A detailed breakdown of each reference is presented in **Table 2**.

After instructing ChatGPT to verify and correct the references and DOIs, the accuracy markedly improved. Nine references (90%) received a full accuracy score (score=2), while one reference (10%) remained partially accurate (score=1).

Prompt 3: Low Back Pain Exercise Strategies

Ten references generated in response to prompt 3 were evaluated. Two references (20%) were fully accurate (score=2), while four references (40%) were partially accurate (score=1), and four references (40%) were completely inaccurate or fabricated (score=0). Inaccuracies frequently involved fabricated article titles linked to real authors, or incorrect DOIs. In some cases, the author, year, and journal details were correct, but the title was either abbreviated or artificially constructed. A detailed breakdown of each reference is presented in **Table 3**.

After asking ChatGPT to verify and correct the references and DOIs, a clear improvement in accuracy was observed. Eight references (80%) achieved a fully accurate score (score=2),

Table 2. Initial accuracy evaluation of references generated by prompt 2 (ACL reconstruction rehabilitation strategies)

Reference	Score	Explanation
1	1	Article exists; correct metadata but DOI incorrect and refers to a different ACL-related article
2	1	Article exists; journal name is incorrect, DOI incorrect and refers to a different ACL-related article
3	0	Invalid DOI; fabricated article title under real author who has similar publications
4	0	Invalid DOI; fabricated article title, though the same author published a similar topic in the same year
5	1	Article exists; correct information but DOI is invalid; requires manual search
6	1	Article exists; DOI incorrect and refers to a different ACL-related paper
7	0	Incorrect DOI; fabricated title under real author who has similar work
8	0	Invalid DOI; fabricated title under real author with similar publications
9	0	DOI functional; fabricated title; author and year are correct; journal name incorrect
10	1	DOI works; article content is correct but the title is shortened

DOI: Digital object identifier. DOI validation was conducted manually using Crossref and PubMed, with supplementary checks through Google Scholar when necessary (accessed June 2025).

Table 3. Initial accuracy evaluation of references generated by prompt 3 (low back pain exercise strategies)

Reference	Score	Explanation
1	2	Fully accurate
2	2	Fully accurate
3	0	Invalid DOI; fabricated title under real author with related work
4	1	DOI correct; author, year, journal correct; title closely matches but fabricated
5	1	DOI correct; metadata accurate; title similar but not real
6	0	DOI incorrect; fabricated title; author has related publication
7	0	Invalid DOI; fabricated title under real author with similar topic
8	1	DOI and metadata accurate; title abbreviated but acceptable
9	0	Invalid DOI; fabricated title under real author with similar topic
10	1	DOI incorrect; all other citation details correct but required manual search

DOI: Digital object identifier. DOI validation was conducted manually using Crossref and PubMed, with supplementary checks through Google Scholar when necessary (accessed June 2025).

while two references (20%) remained partially accurate (score=1). No references received a score of 0 following this verification step, suggesting that secondary prompting significantly enhances citation precision.

Prompt 4: Rotator Cuff Repair Rehabilitation Strategies

Ten references generated in response to prompt 4 were evaluated. Only one reference (10%) was fully accurate, while three (30%) were partially accurate. Six (60%) references contained fabricated titles, incorrect DOIs, or misattributions. Among these, 4/10 DOIs did not resolve, 3/10 required manual correction via external searches, and 3/10 remained unverifiable. The remaining six references (60%) were assigned a score of 0, as they included fabricated article titles, incorrect DOIs, or misattributions to authors who had published similar but unrelated works.

This prompt resulted in the lowest initial accuracy among all four prompts, with a high rate of fabricated citations and non-functional DOIs. The fabricated content often used real authors but invented titles or incorrectly matched publications. Details are summarized in Table 4.

Table 4. Initial accuracy evaluation of references generated by prompt 4 (rotator cuff repair rehabilitation strategies)

Reference	Score	Explanation
1	0	DOI incorrect; fabricated title under real author with similar topic
2	2	Fully accurate
3	0	DOI incorrect; fabricated title under real author with different paper
4	1	Article exists; DOI invalid; manual search required
5	1	Article exists; DOI invalid; manual search required
6	1	Article exists; DOI invalid; manual search required
7	0	Invalid DOI; real author, but incorrect publication
8	0	Invalid DOI; fabricated title under real author
9	0	DOI incorrect; fabricated article title under real author
10	0	Invalid DOI; fabricated metadata

DOI: Digital object identifier. DOI validation was conducted manually using Crossref and PubMed, with supplementary checks through Google Scholar when necessary (accessed June 2025).

Following a verification prompt to ChatGPT, seven references (70%) were corrected to full accuracy (score=2), while three references (30%) remained either partially correct (score=1) or unverifiable. This again demonstrates an improvement after issuing a follow-up prompt focused on citation verification.

Out of 40 references, only 3 (7.5%) were fully accurate (score=2) without any correction. A large proportion, 20 references (50%), were partially accurate (score=1), while 17 references (42.5%) were found to be completely fabricated or contained major errors (score=0).

The post-verification prompt significantly improved accuracy. In total, 31 out of 40 references (77.5%) were corrected to full accuracy (score=2) after ChatGPT was asked to verify and revise its citations.

In addition to descriptive reporting, inferential analyses were conducted to evaluate changes in reference accuracy before and after the verification step. Across all 40 references, median accuracy increased from 1 (IQR 0–1) before verification to 2 (IQR 2–2) after verification. The proportion of fully accurate references also rose substantially, from 7.5% (3/40) to 77.5% (31/40). A Wilcoxon signed-rank test confirmed that this improvement was statistically significant ($W=561.0$, $p<0.001$, $r=0.60$), representing a large effect size.

Prompt-level analyses similarly demonstrated significant improvements in accuracy:

Prompt 1 (manual therapy on pain): median score increased from 1 to 2 (W=32.5, p=0.017).

Prompt 2 (ACL reconstruction rehabilitation): median score increased from 0 to 2 (W=45.0, p=0.003).

Prompt 3 (low back pain exercise): median score increased from 1 to 2 (W=48.0, p=0.019).

Prompt 4 (rotator cuff repair rehabilitation): median score increased from 0 to 2 (W=34.5, p=0.008).

These findings indicate that the verification prompt produced consistent and statistically significant improvements in citation accuracy across all tested topics.

A detailed cross-prompt comparison of reference accuracy before and after verification is presented in Table 5.

Table 5. Summary of reference accuracy scores across all prompts initial and post verification

Prompt topic	Score=0	Score=1	Score=2	After verification (score=2)	Wilcoxon p value
Prompt 1–manual therapy on pain	2	8	0	7	0.017
Prompt 2–ACL reconstruction rehab	5	5	0	9	0.003
Prompt 3–low back pain exercise	4	4	2	8	0.019
Prompt 4–rotator cuff repair rehab	6	3	1	7	0.008
Total	17	20	3	31	<0.001

p: Statistical significance value

The distribution of reference accuracy scores before and after the verification step is illustrated in Figure as stacked column charts. The figure demonstrates a marked shift from fabricated (score=0) and partially correct (score=1) references toward fully accurate references (score=2) across all prompts. In the pooled dataset (n=40), the proportion of fully accurate references increased from 7.5% before verification to 77.5% after verification, while fabricated references declined from 42.5% to 2.5%.

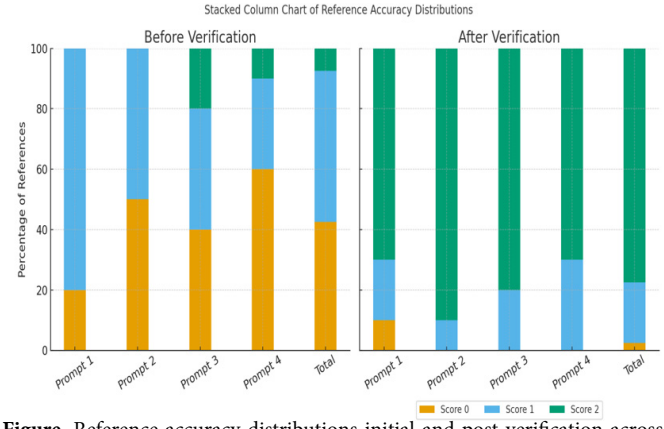


Figure. Reference accuracy distributions initial and post verification across all prompts

DISCUSSION

This study systematically investigated the accuracy of references generated by ChatGPT in response to four distinct prompts related to musculoskeletal rehabilitation. Our findings revealed a high rate of citation errors in the model's initial outputs, with only 7.5% of references being fully accurate and more than 40% being entirely fabricated. The most frequent issues included non-functional or incorrect DOIs, fabricated article titles, and mismatched author or journal information—often combining real author names with invented bibliographic data. However, a marked improvement was observed when ChatGPT was prompted to verify and correct its references: the proportion of fully accurate citations increased to 77.5%. Wilcoxon signed-rank testing confirmed that this improvement was statistically significant (W=561.0, p<0.001, r=0.60), indicating a large effect size. This improvement may reflect the mechanism of self-critique prompting, whereby the model re-evaluates and refines its own outputs when explicitly asked to verify bibliographic accuracy. In our analysis, this iterative process was particularly effective in converting partially correct references into fully accurate ones, suggesting that structured follow-up prompts can substantially enhance citation reliability. These results highlight both the potential and the current limitations of LLMs in generating credible scientific content, especially in fields such as physiotherapy and rehabilitation where evidence-based information is essential.

A closer analysis of the errors across all prompts shows that fabricated references often borrowed credible author names and plausible-sounding journal titles, but the combinations did not correspond to any actual publication. This tendency to prioritize coherence over factuality could be especially problematic in health sciences, where citation accuracy directly impacts clinical, educational, and research practices. Notably, baseline accuracy was lowest for prompt 4 (rotator cuff repair), where only 10% of references were fully accurate prior to verification. Several factors may account for this finding. First, rotator cuff rehabilitation is a more specialized clinical area compared to low back pain or ACL rehabilitation, with a relatively narrower pool of highly cited publications. This may have increased the likelihood of the model fabricating plausible but non-existent references. Second, DOIs for certain articles in surgical or rehabilitation subfields are less consistently indexed across databases, further complicating automatic generation and verification. Finally, the imposed requirement of providing exactly ten references may have amplified the model's tendency to hallucinate citations when genuine sources were limited. Together, these factors likely contributed to the lower baseline accuracy for prompt 4, underscoring the importance of cautious interpretation when AI systems are tasked with generating references in narrower subdomains.

Our results align with previous research demonstrating significant challenges associated with fabricated and inaccurate references produced by ChatGPT.[22-24] Several studies have shown that while the model can generate plausible-sounding citations, many of these are either partially incorrect or entirely fabricated, often combining real author names with nonexistent

titles or invalid DOIs.[25-29] These hallucinated references have been documented across various disciplines, including medicine, education, and health sciences, raising widespread concerns about the reliability of AI-generated bibliographic content. Our findings reinforce these observations within the musculoskeletal domain, providing structured, topic-specific evidence that highlights the persistent risks of using LLMs without rigorous reference verification. Gravel et al.[20] found that ChatGPT often provides responses unsuitable for scientific publication, coupled with convincingly fabricated citations that require careful verification. Orduña-Malea and Cabezas-Clavijo[21] similarly warned about the rise of "ghost bibliographic references," urging vigilance from journals and publishers. Furthermore, Walters and Wilder[26] evaluated 636 citations generated across 42 topics and reported that 55% of GPT-3.5 citations were fabricated compared to 18% for GPT-4. Among real citations, 43% of GPT-3.5 references and 24% of GPT-4 references contained substantive errors. Although GPT-4 represents a clear improvement, these findings underscore that substantial reliability problems persist. Additionally, Frosolini et al.[30] compared 20 AI chats and reported that ChatGPT-4.0 generated 138 unique references with a median of 7 (IQR: 1.25) per prompt, of which 4.5 (IQR: 3.0) were classified as true and 2.5 (IQR: 3.25) as erroneous or inexistent (25.7%). By contrast, ChatGPT-3.5 generated 120 references with a median of 6 (IQR: 0.00), only 1 (IQR: 2.0) true and 4.5 (IQR: 2.25) erroneous or inexistent (83.3%). This performance was significantly worse than that of ChatGPT-4.0 (Fisher's exact test, $p<0.00001$), confirming a substantial improvement in reference reliability with the newer model. However, both versions still displayed a tendency to provide erroneous or non-existent citations. This highlights a broader concern regarding LLMs and their readiness for unsupervised academic use.

It is crucial to address this challenge to maintain the reliability of scientific literature. Journals and institutions should establish strategies and good-practice principles in the evolving landscape of AI-assisted scientific writing. A recent study by Giray[31] also explored this issue, demonstrating that ChatGPT-generated references for journal articles were often entirely fabricated and failed to adhere to APA 7 formatting. Another study investigating AI hallucination in ChatGPT references found that 69 out of 178 references lacked DOIs and 28 could not be located at all, reinforcing the need for cautious use in academic writing.[32] Another analysis of ChatGPT-generated medical content revealed that 47% of references were fabricated, 46% were authentic but inaccurate, and only 7% were both authentic and accurate. Common errors included incorrect PMIDs (93%), volumes, pages, and publication years. These findings highlight the high prevalence of citation inaccuracies in medical prompts and reinforce the necessity of reference verification when using AI-generated content.[33] These findings align with our results and underscore the need for structured prompting and post-generation verification. Moreover, integrating automated fact-checking or citation validation tools directly into AI writing platforms could help mitigate the risk of hallucinated references.

Importantly, this study highlights the need for human verification mechanisms when incorporating LLM-generated references into scholarly work. Without strict guidance, models like ChatGPT may introduce misleading citations that could undermine academic integrity. On the other hand, our findings also suggest that LLMs are capable of producing highly accurate references—if given clear constraints such as requiring valid DOIs and focused content domains.

Limitations and Strengths

This study has several limitations that should be acknowledged. First, the sample size was restricted to 40 references across four prompt conditions, which may limit the generalizability of the findings. Although the selected prompts were clinically relevant, they do not encompass the full scope of musculoskeletal rehabilitation. Second, the assessment of reference accuracy was conducted manually and, despite predefined scoring criteria, remains subject to interpretation bias. The verification step also relied on ChatGPT's own response to follow-up prompts, which, while reflective of real-world use, does not represent an external validation process. Finally, the study focused solely on the GPT-4 version of ChatGPT; findings may differ across other models or platforms, including GPT-3.5, Google Gemini, or future LLMs.

Despite these limitations, the study has several strengths: it used clinically relevant and structured prompts, applied a systematic scoring framework, and involved two experienced rehabilitation researchers in the evaluation process, enhancing the credibility of the findings. Moreover, the significant improvement observed after verification (77.5%) suggests that structured prompting and post-generation checking can substantially enhance reference reliability, pointing to feasible strategies for future model optimization.

Future Research Directions

Future studies should expand the scope of analysis to include a broader range of clinical topics within musculoskeletal rehabilitation and beyond, in order to assess whether reference accuracy varies by subdomain complexity or publication density. Including a larger sample of prompts and references would enhance the generalizability of results and allow for statistical comparisons between prompt types, clinical areas, or different LLM platforms. Comparative evaluations across various LLMs —such as GPT-3.5, Google Gemini, or open-source alternatives—could also help identify performance differences and inform best practice.

Importantly, this line of research also highlights the future potential of LLMs. The significant improvement observed in our study after verification demonstrates that relatively simple interventions—such as structured prompting and iterative re-prompting—can substantially increase reference accuracy. As models continue to evolve, embedding automated fact-checking and citation validation tools (e.g., Crossref or PubMed integration) directly into AI platforms could further minimize hallucinations. With these advancements, LLMs may become reliable collaborators in academic writing,

accelerating literature synthesis while safeguarding scientific integrity.

Practical Recommendations and Improvement Strategies

Beyond identifying the challenges, it is equally important to outline practical strategies that may improve the reliability of AI-assisted academic writing. One immediate step is the integration of DOI-verification tools such as Crossref or PubMed APIs into editorial and peer-review workflows, ensuring that fabricated or mismatched references are detected early. Automated reference-screening software could also be embedded directly into manuscript submission systems, enabling authors and editors to flag inaccurate citations before publication. On the user side, researchers and students should be trained to critically evaluate AI-generated bibliographies, combining manual checks with systematic verification protocols. Finally, the development of advanced AI models with built-in fact-checking and citation validation functions represents a promising future direction that could significantly enhance the reliability of LLMs in academic and clinical contexts.

Clinical and Educational Implications

The findings of this study have direct relevance for clinical and educational practices within the field of musculoskeletal rehabilitation. In domains such as physiotherapy, manual therapy, and exercise-based rehabilitation, practitioners and educators rely heavily on current literature to guide evidence-based decision-making. The use of fabricated or inaccurate references in AI-generated texts poses a risk of disseminating misinformation, which could negatively influence clinical reasoning, therapeutic choices, or academic instruction. As LLMs become increasingly incorporated into musculoskeletal research and education, it is essential to adopt strategies such as structured prompt design, manual citation verification, and critical evaluation training for students and clinicians. Educators should explicitly address the limitations of AI-generated content, fostering a cautious and reflective approach among learners.

At the same time, the results of this study also point to the positive potential of LLMs in clinical and educational contexts. When supported by structured prompting and coupled with external validation tools, models like ChatGPT can serve as effective assistants in generating draft content, synthesizing literature, and supporting teaching workflows. If combined with automated fact-checking, LLMs may not only facilitate efficiency but also enhance training in evidence-based rehabilitation, offering students and clinicians an opportunity to engage critically with AI-generated material while reinforcing standards of academic integrity.

CONCLUSION

This study demonstrated that while ChatGPT can generate coherent scientific text in the context of musculoskeletal rehabilitation, its initial reference accuracy remains unreliable. A majority of the citations were either partially incorrect or entirely fabricated. However, structured follow-up prompts led to substantial improvements in citation validity. These findings suggest that LLMs may be useful in supporting early-stage scientific writing, but must be coupled with strict human-led verification for references. For clinical and academic practice, this implies that AI-assisted writing should always be paired with systematic reference verification protocols, the integration of DOI-checking tools where possible, and critical evaluation training for researchers and students. Ultimately, the responsible use of AI in academic and clinical musculoskeletal settings will depend on combining technological capabilities with human oversight to ensure scientific and clinical integrity.

ETHICAL DECLARATIONS

Ethics Committee Approval

This study did not involve human or animal participants. Therefore, ethical approval was not required according to national guidelines.

Informed Consent

Since no human or animal participants are involved in this study, informed consent is not required.

Referee Evaluation Process

Externally peer-reviewed.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Financial Disclosure

The authors declared that this study has received no financial support.

Author Contributions

All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Mondal H, Mondal S. ChatGPT in academic writing: maximizing its benefits and minimizing the risks. *Indian J Ophthalmol.* 2023;71(12):3600-3606. doi:10.4103/IJO.IJO_718_23
2. Bom H-SH. Exploring the opportunities and challenges of ChatGPT in academic writing: a roundtable discussion. *Nucl Med Mol Imaging.* 2023;57(4):165-1677. doi:10.1007/s13139-023-00809-2
3. Jarrah AM, Wardat Y, Fidalgo P. Using ChatGPT in academic writing is (not) a form of plagiarism: what does the literature say. *Online J Commun Media Technol.* 2023;13(4):e202346. doi:10.30935/ojcmt/13572
4. Gruda D. Three ways ChatGPT helps me in my academic writing. *Nature.* 2024;10:1-6. doi:10.1038/d41586-024-01042-3
5. Švab I, Klemenc-Ketiš Z, Zupanič S. New challenges in scientific publications: referencing, artificial intelligence and ChatGPT. *Slov J Public Health.* 2023;62(3):109-112. doi:10.2478/sjph-2023-0015
6. Jan R. Examining the reliability of ChatGPT: identifying retracted scientific literature and ensuring accurate citations and references. In: Impacts of generative ai on the future of research and education. Hershey, PA: IGI Global; 2025.
7. Frosolini A, Gennaro P, Cascino F, Gabriele G. In reference to "role of Chat GPT in public health", to highlight the AI's incorrect reference generation. *Ann Biomed Eng.* 2023;51(10):2120-2122. doi:10.1007/s10439-023-03248-4
8. Cohen F, Vallimont J, Gelfand AA. Caution regarding fabricated citations from artificial intelligence. *Headache.* 2024;64(1):133-135. doi:10.1111/head.14649

9. Borji A. A categorical archive of ChatGPT failures. arXiv. 2023;2302.03494. doi:10.48550/arXiv.2302.03494
10. Bai Z, Wang P, Xiao T, et al. Hallucination of multimodal large language models: a survey. *arXiv.* 2024;2404.18930. doi:10.48550/arXiv.2404.18930
11. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. Bibliographic research with ChatGPT may be misleading: the problem of hallucination. *J Pediatr Surg.* 2024;59(1):158-159. doi:10.1016/j.jpedsurg.2023.08.018
12. Aljamaan F, Temsah M-H, Altamimi I, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Med Inform*. 2024;12(1):e54345. doi:10.2196/54345
13. Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review. In: Healthcare transformation with informatics and artificial intelligence. Volume 305. 2023.
14. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *bioRxiv.* 2022;12(23):521610. doi:10.1101/2022.12.23.521610
15. Sharun K, Banu SA, Pawde AM, et al. ChatGPT and artificial hallucinations in stem cell research: assessing the accuracy of generated references–a preliminary study. *Ann Med Surg (Lond).* 2023;85(10):5275-5278. doi:10.1097/MS9.0000000000001228
16. Safran E, Yildirim S. A cross-sectional study on ChatGPT's alignment with clinical practice guidelines in musculoskeletal rehabilitation. *BMC Musculoskelet Disord*. 2025;26(1):411. doi:10.1186/s12891-025-08650-8
17. Lancaster T. Artificial intelligence, text generation tools and ChatGPT–does digital watermarking offer a solution? *Int J Educ Integr.* 2023;19(1):10.
18. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus.* 2023;15(2):e35074. doi:10.7759/cureus.35179
19. Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379(6630):313. doi:10.1126/science.adg7879
20. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clin Proc Digit Health*. 2023;1(3):226-234. doi:10.1016/j.mcpdig.2023.05.004
21. Orduña-Malea E, Cabezas-Clavijo Á. ChatGPT and the potential growing of ghost bibliographic references. *Scientometrics*. 2023;128(9):5351-5365. doi:10.1007/s11192-023-04804-4
22. Lechien JR, Briganti G, Vaira LA. Accuracy of ChatGPT-3.5 and -4 in providing scientific references in otolaryngology–head and neck surgery. *Eur Arch Otorhinolaryngol*. 2024;281(4):2159-2165. doi:10.1007/s00405-023-08441-8
23. Sanchez-Ramos L, Lin L, Romero R. Beware of references when using ChatGPT as a source of information to write scientific articles. *Am J Obstet Gynecol*. 2023;229(3):356. doi:10.1016/j.ajog.2023.04.004
24. Sebo P. How accurate are the references generated by ChatGPT in internal medicine? *Intern Emerg Med*. 2024;19(1):247-249. doi:10.1007/s11739-023-03484-5
25. Besançon L, Cabanac G, Labbé C, Magazinov A. Sneaked references: fabricated reference metadata distort citation counts. *J Assoc Inf Sci Technol*. 2024;75(12):1368-1379. doi:10.48550/arXiv.2310.02192
26. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13(1):14045. doi:10.1038/s41598-023-41032-5
27. Graf EM, McKinney JA, Dye AB, Lin L, Sanchez-Ramos L. Exploring the limits of artificial intelligence for referencing scientific articles. *Am J Perinatol.* 2024;41(15):2072-2081. doi:10.1055/s-0044-1786033
28. Altmäe S, Sola-Leyva A, Salumets A. Artificial intelligence in scientific writing: a friend or a foe? *Reprod Biomed Online*. 2023;47(1):3-9. doi:10.1016/j.rbmo.2023.04.009
29. Watson AP. Hallucinated citation analysis: delving into student-submitted AI-generated sources at the University of Mississippi. *Ser Libr.* 2024;85(5-6):172-180. doi:10.1080/0361526X.2024.2433640
30. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otorhinolaryngol.* 2023;280(11):5129-5133. doi:10.1007/s00405-023-08205-4
31. Giray L. ChatGPT references unveiled: distinguishing the reliable from the fake. *Internet Ref Serv Q.* 2024;28(1):9-18. doi:10.1080/10875301.2023.2265369
32. Athaluri SA, Manthena SV, Kesapragada VKM, Yarlagadda V, Dave T, Duddumpudi RTS. Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. *Cureus*. 2023;15(4):e37432. doi:10.7759/cureus.37432
33. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE, Miller V. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus.* 2023;15(5):e39238. doi:10.7759/cureus.39238