scientific reports

Check for updates

OPEN

Evaluating ChatGPT's ability to simplify scientific abstracts for clinicians and the public

Esra Dogru-Huzmeli[1]✉, Sarah Moore-Vasram[2,3], Chetan Phadke[2,4], Erfan Shafiee[5] & Shabbir Amanullah[6]

This study evaluated ChatGPT's ability to simplify scientific abstracts for both public and clinician use. Ten questions were developed to assess ChatGPT's ability to simplify scientific abstracts and improve their readability for both the public and clinicians. These questions were applied to 43 abstracts. The abstracts were selected through a convenience sample from Google Scholar by four interdisciplinary reviewers from physiotherapy, occupational therapy, and nursing backgrounds. Each abstract was summarized by ChatGPT on two separate occasions. These summaries were then reviewed independently by two different reviewers. Flesch Reading Ease scores were calculated for each summary and original abstract. A subgroup analysis explored differences in accuracy, clarity, and consistency across various study designs. ChatGPT's summaries scored higher on the Flesch Reading Ease test than the original abstracts in 31 out of 43 papers, showing a significant improvement in readability ($p = 0.005$). Systematic reviews and meta-analyses consistently received higher scores for accuracy, clarity, and consistency, while clinical trials scored lower across these parameters. Despite its strengths, ChatGPT showed limitations in "Hallucination presence" and "Technical terms usage," scoring below 7 out of 10. Hallucination rates varied by study type, with case reports having the lowest scores. Reviewer agreement across parameters demonstrated consistency in evaluations. ChatGPT shows promise for translating knowledge in clinical settings, helping to make scientific research more accessible to non-experts. However, its tendency toward hallucinations and technical jargon requires careful review by clinicians, patients, and caregivers. Further research is needed to assess its reliability and safety for broader use in healthcare communication.

Keywords ChatGPT, Flesch reading ease score, Hallucination presence, Technical terms, Healthcare dissemination

Background

ChatGPT is an artificial intelligence model developed by OpenAI that uses advanced natural language processing methods. It can understand and generate human-like text based on user input. Whether used for retrieving information, casual conversation, or solving problems[1] ChatGPT has become a versatile tool with broad applications in healthcare, including medical translation and education.

This technology shows potential to improve diagnostic accuracy, respond to patient questions, and personalize healthcare advice such as lifestyle recommendations[2–4].

However, since AI is still relatively new in the health sciences, its safety, effectiveness, and reliability for daily clinical use are still being evaluated.

Current research emphasizes the need to address these challenges while leveraging the benefits of ChatGPT in medical settings[2]. Concerns have been raised about the validity and accuracy of AI-generated summaries, especially when the audience includes patients and caregivers. The use of complex technical language makes it difficult to communicate key information clearly. This issue also affects healthcare professionals, who face increasing challenges in interpreting complex data.

[1]Faculty of Health Science, Physiotherapy and Rehabilitation Department, Istanbul Gedik University, Istanbul, Türkiye. [2]Providence Care Centre, Kingston, ON, Canada. [3]School of Nursing, Queen's University, Kingston, ON, Canada. [4]School of Rehabilitation Therapy, Queen's University, Kingston, ON, Canada. [5]Bruyère Health, Ottawa, ON, Canada. [6]University of Western Ontario, London, ON, Canada. ✉email: esradogru001@gmail.com

The aim of this study is to evaluate ChatGPT's ability to simplify scientific abstracts and improve their readability for both the public and clinicians.

Methods

Protocol

Ten questions were developed to evaluate ChatGPT's ability to simplify scientific abstracts for improved readability by both the public and clinicians. The primary investigator generated the questions based on a preliminary literature review and these were reviewed by the co-authors with consideration given to real world feasibility and application. The selection of coauthors was designed to represent a variety of health care professionals that included physiotherapy, occupational therapy, nursing, and medicine. One coauthor, a medical expert in ChatGPT, was excluded from the abstract review process to ensure impartiality.

Consensus was reached for the final 10 questions, and the ten questions were scored on a scale from 0 to 10, with 0 representing the worst and 10 representing the best result.

Reviewers were advised to select research abstracts that relate to their work in the hospital and a sample of 43 research abstracts were selected. Google Scholar was chosen as the search engine due to its global accessibility, while authors accepted that there maybe an inherent bias in the coding algorithm that could skew search results.

Four reviewers independently copied and pasted their selected abstracts into ChatGPT for summary on two separate occasions (Fig. 1). These papers were distributed among the four reviewers for initial assessment in an excel spreadsheet. Subsequently, each paper underwent a secondary evaluation by different reviewers, resulting in a total of two evaluations per paper by two different reviewers. Reviewers were blinded to the results and

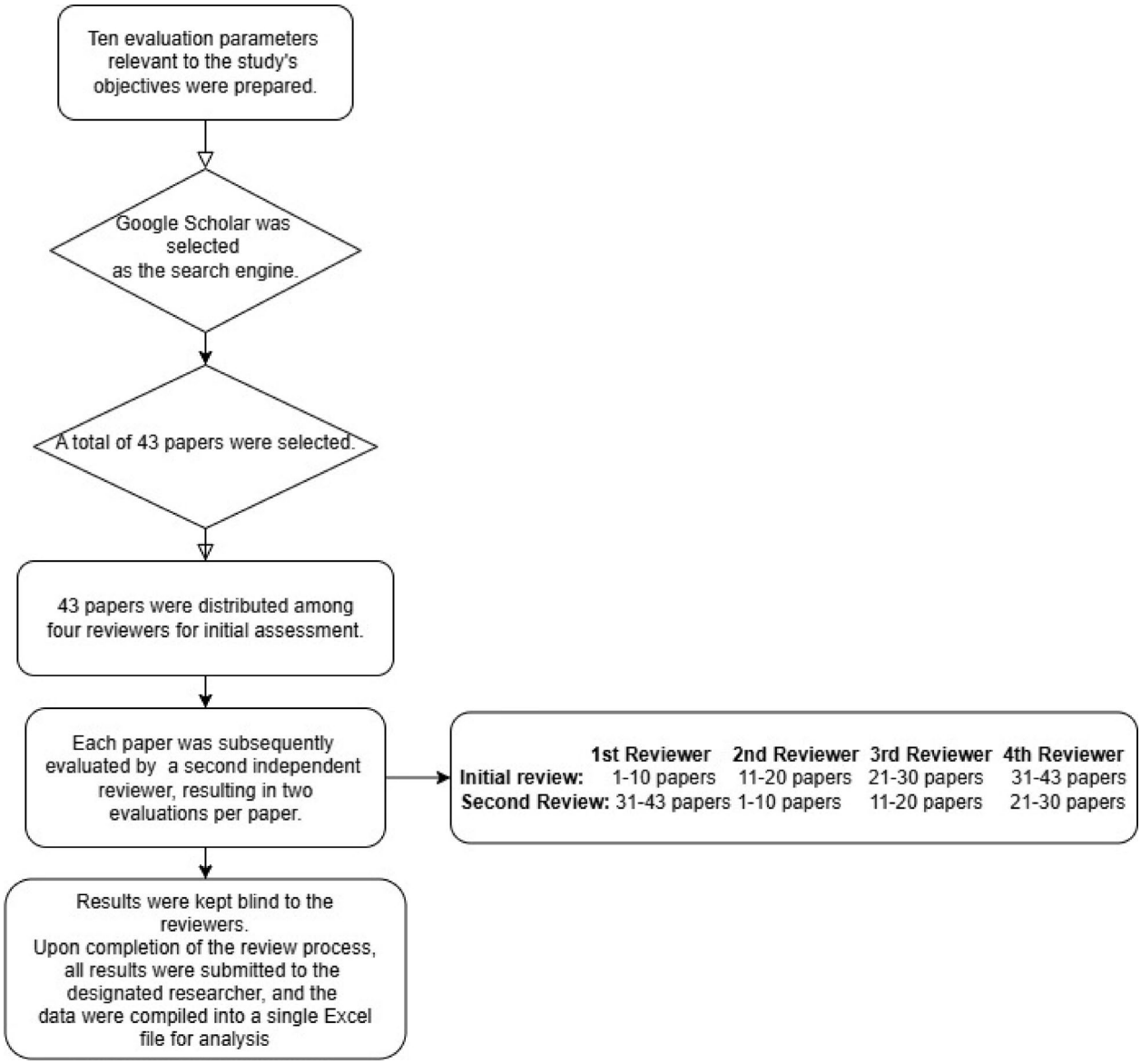


Fig. 1. Flowchart of the study.

upon completion of the review process, all results were submitted to the designated researcher, and the data was compiled into a single Excel file for analysis.

Google scholar search

To identify relevant abstracts, the following keywords were used in Google Scholar: "stroke physiotherapy and rehabilitation", "nursing", "occupational therapy", and "physical therapy". Studies from the last five years were selected, and the article type was set to "any type" on Google Scholar. Publications were then randomly selected from the search results, ensuring diversity in publication types. In our study, we aimed to include various types of studies such as meta-analyses, reviews, case studies, qualitative studies, cohort studies, randomized controlled studies, and non-randomized controlled studies. This diversity in study designs was intentional, allowing assessment of ChatGPT's ability to interpret different research methodologies.

Inclusion criteria for paper selection:

- Papers related to occupational therapy, physical therapy or nursing.
- Study designs included meta-analyses, reviews, case studies, qualitative studies, cohort studies, randomized controlled trials, and non-randomized controlled trials.
- Studies published within the last 15 years.

Forty-three papers were selected for review[5–47].

ChatGPT task

Reviewers used a standardized prompt: "*Can you please provide a layman's summary of the main findings and implications from this scientific publication? Make sure to avoid technical jargon and focus on explaining the key points for a general audience*". This prompt was designed to encourage simplified language while retaining key findings. The title and abstract of the study were then added by placing a colon right after the question. ChatGPT was blinded to the authors' names and affiliations in the papers being reviewed by deleting those details.

Evaluation process

Four reviewers were involved in the ChatGPT abstract summary evaluation. 3 reviewers selected 10 articles from their field of expertise and completed the first evaluation. 1 reviewer selected 13 articles and completed 13 evaluations. Each summary that was provided by ChatGPT was evaluated by 2 reviewers, and their scores were averaged. Four reviewers conducted cross-evaluations to ensure homogeneity. The reviewers were blinded to the evaluation results of the other reviewers to mitigate bias. Each publication was evaluated by two different reviewers.

We scored each parameter on a scale of 0–10. Getting a high score meant that ChatGPT was successful and generated a result that best-fit the evaluated parameter. We defined success as a value of 7 out of 10. This threshold was an arbitrary yet pragmatic cut-off determined by the principal investigator to reflect high performance on a 0–10 scale. While heuristic in nature, it aligns with commonly accepted evaluative standards in academic and professional contexts, where a score of ≥70% is typically regarded as satisfactory or proficient. To ensure the accuracy of the summaries, each ChatGPT-generated abstract was compared to the original version. Abstracts were broken down into key components (Purpose, Methods, Results, Discussion, and Conclusions), and ChatGPT's outputs were evaluated for completeness and factual accuracy using an objective content matching method.

Evaluated parameters

1. Rating the best-fit for each part of the abstract (purpose, method, results, discussion) on a 0–10 Scale. This parameter evaluated the extent to which ChatGPT delivered the main result that the original abstract conveyed, ranging from 0 (worst) to 10 (best). A score of 0 indicates an incomplete result with a significant amount of missing information. A score of 10 indicated a complete and comprehensive summary of the information.
2. Accuracy of Content of Purpose-Method-Results-Discussion-Conclusions (0–10 Scale): This parameter evaluates the accuracy of the content presented in the ChatGPT text compared to the original abstract. A score of 0 indicates complete inaccuracy, while a score of 10 indicated complete accuracy in alignment with the main abstract.
3. Hallucination Presence Assessment (0–10 Scale): This parameter evaluated if ChatGPT gave results that did not match the original abstract. A score of 0 indicates the highest volume of experiencing imaginary results that did not align with the abstract, while a score of 10 indicated the lowest presence.
4. Technical Terms Usage Assessment (0–10 Scale): This parameter evaluated the presence of words or terminology that may be difficult for the general public to understand. A score of 0 indicates the extensive use of complex language that may not be suitable for non-healthcare professionals, while a score of 10 indicated the absence of complex language.
5. Consistency Assessment (0–10 Scale): This parameter evaluates the degree of consistency in results of ChatGPT when the same question is asked at different times by different reviewers. All the abstracts that were extracted from Google Scholar, were recorded in an excel spreadsheet. The ChatGPT abstract summary results at time 1 and time 2 results were recorded in the same excel document. The reviewers' assessments were made according to the recorded results at time 1. For the consistency assessment, the same abstract and question were submitted to ChatGPT a second time. So, essentially, the same abstract was asked of ChatGPT twice at different times. The results given by ChatGPT to two different reviewers at two different times were compared and their compatibility with each other was evaluated. This parameter was scored on

a scale of 0–10 and the average of the scores was used to evaluate Chat GPT's response. The task date was recorded as well.

6. Clarity Assessment (0–10 Scale): This parameter evaluates ChatGPT's ability in simplifying complex scientific language into clear and easily understandable terms. Factors such as coherence, conciseness, and overall readability were considered when assigning a score on the scale from 0 (poor clarity) to 10 (excellent clarity).
7. ChatGPT Patient/ Community Recommendation: We asked if the reviewer would recommend ChatGPT to patient and community members. This was rated on a scale from 0 (never) to 10 (always).
8. ChatGPT Summary Recommendation: We asked if the reviewer would recommend ChatGPT to their colleagues. This was scored on a scale between 0 (never) and 10 (always).
9. Satisfaction with ChatGPT Summary: We asked the reviewers about their satisfaction level with the abstract summary. This was scored on a scale of 0 (not satisfied) to 10 (completely satisfied).
10. Readability Score: We used Microsoft Word 365's embedded Flesch Reading Ease (FRE) test to measure readability. We used Flesch Reading Ease results for this task[48,49]. Each readability test bases its evaluation on the average number of syllables per word and words per sentence generating a subsequent rating on a 100-point scale[48,49]. Flesch Reading Ease test rates text on a 100-point scale. The higher the score, the easier it is to understand the document. For most standard files, the score preferred is between 60 and 70. We evaluated the main abstract's readability and the readability of the abstract that ChatGPT provided. We copied the summaries into a blank Word document and evaluated their reliability using the embedded FRE test.

To define success, we adopted a threshold score of 7 out of 10 for the 1,2,3,4,5,6,7,8,9 item. This value was initially chosen as a practical and conservative benchmark by the principal investigator. A score of 7 or higher was considered indicative of a high-quality, contextually appropriate response. This threshold is consistent with evaluative norms where scores of 70% or above are typically seen as satisfactory or successful.

Statistical analyses

Analysis was completed using IBM SPSS statistics V 22.0. The descriptive statistics of numerical data were reported as mean and standard deviation (X ± SD), and the descriptive statistics of proportional data were reported as frequencies and percentages (n, %). The normality of distribution was examined with the Shapiro-Wilk test. For FRE readability score analyses we used the Wilcoxon Signed Rank Test. P was defined as ≤ 0.05.

Results

All items except 'Hallucination presence' and 'Technical terms usage' received scores greater than 7 out of 10. Content accuracy yielded the highest score while hallucination presence was the lowest score (see Table 1). We defined a score of 7 out of 10 as the threshold for a "successful" outcome. This threshold was selected by the principal investigator as a practical benchmark, aligning with common evaluative norms in academic and applied settings (e.g., a score ≥ 70% as indicative of acceptable or proficient performance). Our data support this distinction: most criteria have mean scores at or above 7, with two exceptions-hallucination presence (X = 6.01) and technical terms usage (X = 6.95). Notably, hallucination presence is clearly below the threshold and was also identified as an area with qualitative concerns. Technical terms usage, while slightly below 7, remains very close to the threshold, suggesting borderline adequacy. This pattern supports the 7-point cut-off as a meaningful dividing line between stronger and weaker areas of performance. This natural separation reinforces the 7-point threshold as a meaningful dividing line between successful and underperforming outputs.

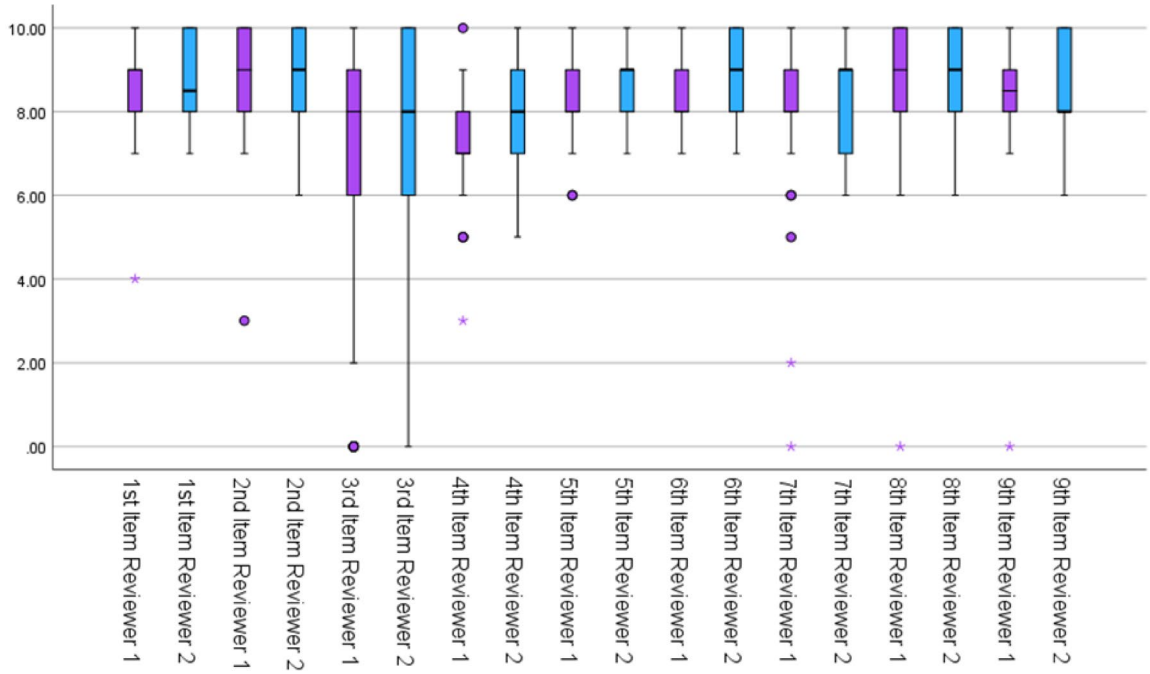
Papers were categorized as clinical trials (qualitative studies, cohort studies, randomized controlled studies, and non-randomized controlled studies), systematic reviews, meta-analyses, and case reports. We observed that, across most parameters, systematic reviews and meta-analyses obtained higher scores, whereas clinical trials received lower scores.

	All Papers X ± SD (0–10 Scale) (*n* = 43)	Clinical Trials X ± SD (0–10 Scale) (*n* = 24)	Systematic Reviews and Meta-analyses X ± SD (0–10 Scale) (*n* = 14)	Case Reports X ± SD (0–10 Scale) (*n* = 5)
1. Rating the Best Fit for each part of the abstract (purpose, method, results, discussion)	7.87 ± 1.73	7.47 ± 2.00	8.57 ± 1.03	7.80 ± 1.60
2. Accuracy of Content	8.08 ± 1.89	7.70 ± 2.13	8.67 ± 1.30	8.20 ± 1.95
3. Hallucination Presence Assessment	6.01 ± 2.07	6.18 ± 2.19	6.03 ± 2.22	5.10 ± 0.54
4. Technical Terms Usage Assessment	6.95 ± 1.59	6.87 ± 1.94	7.25 ± 0.91	6.50 ± 1.32
5. Consistency Assessment	7.75 ± 1.74	7.35 ± 1.99	8.42 ± 0.85	7.80 ± 2.01
6. Clarity Assessment	7.93 ± 1.81	7.50 ± 2.16	8.71 ± 0.64	7.80 ± 1.68
7. Would you recommend ChatGPT to patients and community members?	7.50 ± 1.79	7.35 ± 2.00	7.96 ± 1.56	6.90 ± 1.14
8. Would you recommend ChatGPT as a method for summarizing research in layman terms to a colleague?	7.98 ± 1.96	7.52 ± 2.15	8.78 ± 1.38	8.00 ± 2.00
9. How satisfied are you with the abstract provided by ChatGPT?	7.74 ± 1.71	7.45 ± 1.93	8.42 ± 1.25	7.20 ± 1.35

Table 1. Reviewer ratings of ChatGPT-Generated abstracts across study Types. X ± SD = Mean ± Standard Deviation. Scores for each evaluation parameter ranged from 0 to 10, where 0 indicated the lowest (most negative) and 10 the highest (most positive) outcome.

Flesch Reading Ease			
	X ± SD (min-max)	p	z
Abstract Results (0-100)	14.55 ± 10.58 (0-50.4)	**0.005**	-2.789
ChatGPT Results (0-100)	22.36 ± 12.10 (2.4–51.60)		

Table 2. Flesch reading ease Results. Wilcoxon Signed Rank Test. X ± SD = Mean ± Standard Deviation. Flesch Reading Ease scores range from 0 to 100, with higher scores indicating greater readability.

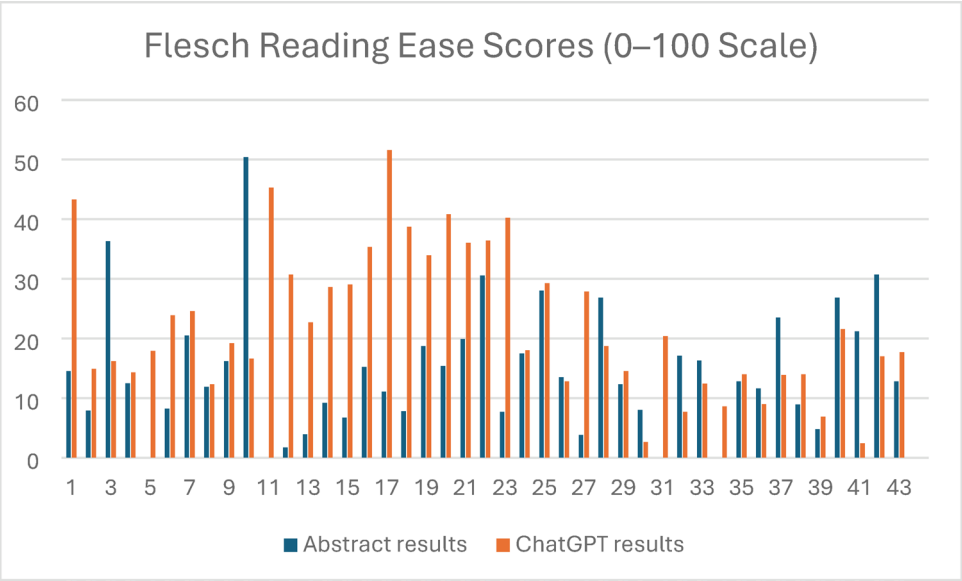


Graph. 1. Comparison of first and second reviewer scores for evaluation parameters (Mean scores on a 0–10 scale) 1st Item: Rating the Best Fit Result, 2nd Item: Accuracy of Content, 3rd Item: Hallucination Presence Assessment, 4th Item: Technical Terms Usage Assessment, 5th Item: Consistency Assessment, 6th Item: Clarity Assessment, 7th Item: Recommending ChatGPT to Patients and Community Members, 8th Item: Recommending ChatGPT to a colleague, 9th Item: Satisfaction Status with the Abstract Provided by ChatGPT. Scores for each evaluation parameter ranged from 0 to 10, where 0 indicated the lowest or most negative result, and 10 indicated the highest or most positive result.

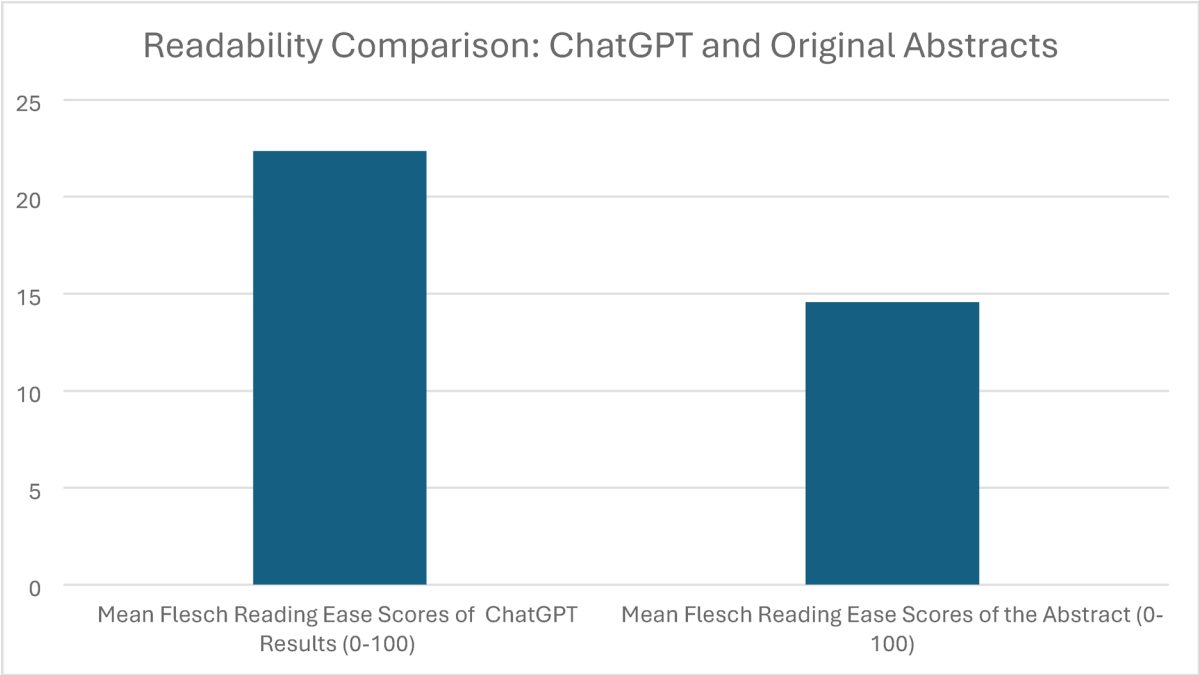
Flesch reading scores were better for the AI summary in 31 out of 43 articles. That test result indicated that the ChatGPT summary was more readable than the original abstract. The ChatGPT summaries were approximately 34.92% [(22.36–14.55 = 7.81), (7.81 × 100)/22.36)] more readable than the original abstracts and this difference was significant ($p = 0.005$), (Table 2). The table displays mean ± standard deviation (X ± SD) of Flesch Reading Ease (FRE) scores for original abstracts and ChatGPT-generated summaries ($n = 43$). FRE scores range from 0 to 100, where higher scores indicate greater readability. The mean FRE score for original abstracts was 14.55 ± 10.58 (range: 0–50.4), while ChatGPT summaries scored 22.36 ± 12.10 (range: 2.4–51.60). The difference was statistically significant ($p = 0.005$).

We found a high level of agreement between the two reviewers of each paper for each evaluated parameter (see Graph 1). In the graph, we displayed the mean scores given by each reviewer for each parameter (9 items). The purple bars represent the mean scores from the first reviewer, while the blue bars represent the mean scores from the second reviewer. Each research abstract included in the study was independently assessed by two different reviewers to enhance objectivity and reliability of the evaluation process. For each of the nine evaluation parameters—such as rating the best fit, accuracy of content, hallucination presence, technical terminology, consistency, clarity, recommendation for patient and clinician use, and overall satisfaction—two separate scores were recorded per paper. These scores were provided independently and without knowledge of the other reviewer's ratings. The final analysis used the mean of the two scores for each parameter to generate the results presented in Table 1.

In Graph 2, titled *"Comparison of Flesch Reading Ease Scores: Abstract vs. ChatGPT Results for Each Paper"*, the data visually contrasts the Flesch Reading Ease (FRE) scores of the original abstracts and their corresponding ChatGPT summaries across each individual paper. The Flesch Reading Ease score is an important measure of



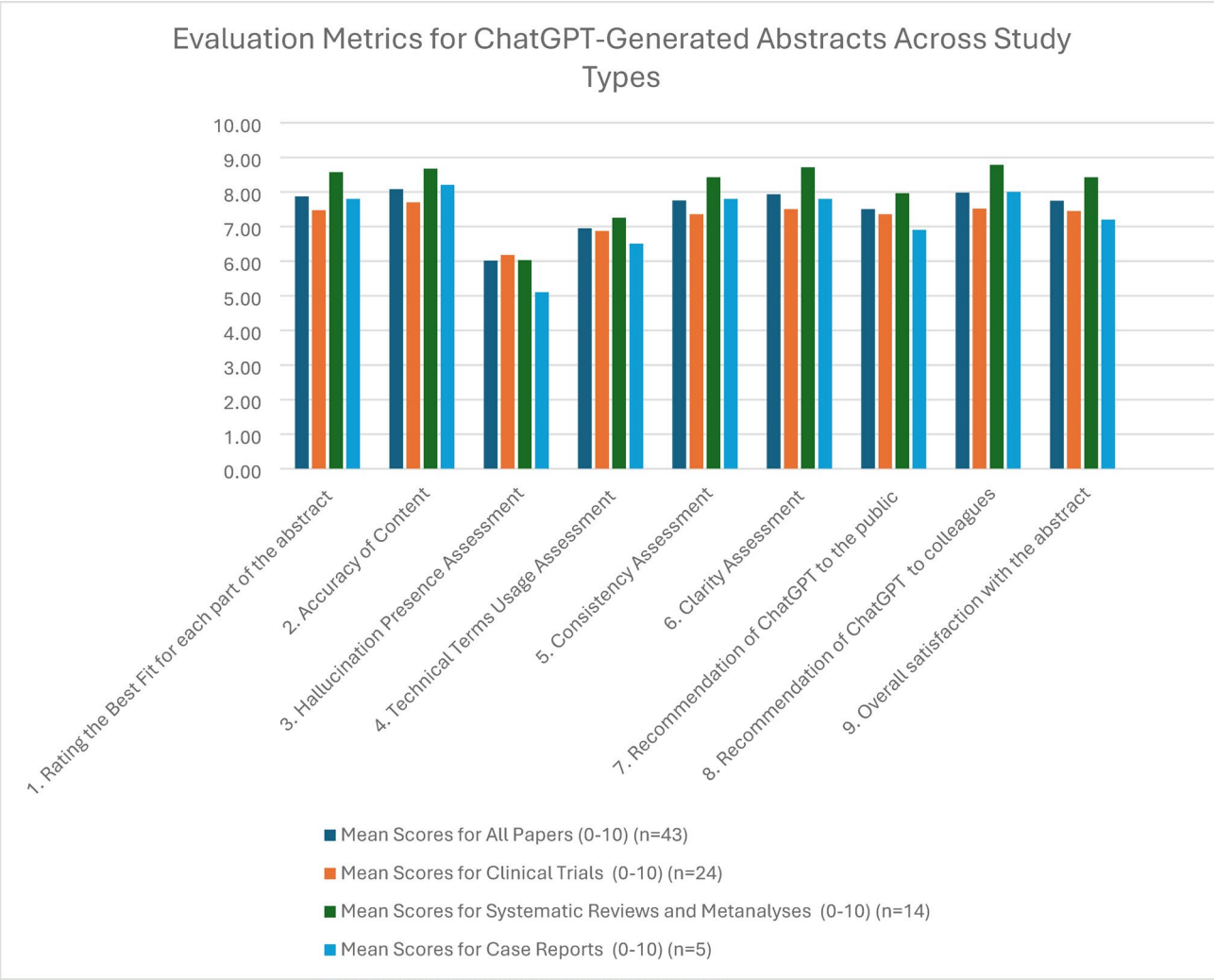
Graph. 2. Comparison of Flesch Reading Ease Scores Between Original and ChatGPT-Generated Abstracts Across 43 Papers. Flesch Reading Ease scores range from 0 to 100, where higher scores indicate greater readability.



Graph. 3. Comparison of Mean Flesch Reading Ease Scores: ChatGPT Summaries vs. Original Abstracts. Flesch Reading Ease scores range from 0 to 100, where higher scores indicate greater readability.

readability, with higher scores indicating easier comprehension. This comparison provides insight into how well ChatGPT can simplify the language of scientific abstracts to make them more accessible to a broader audience, including those without a background in the field. In Graph 3, titled *"Comparison of Mean Scores of Flesch Reading Ease Scores: ChatGPT Summaries vs. Original Abstracts"*, the average Flesch Reading Ease scores for all the papers are compared between the original abstracts and the ChatGPT-generated summaries. This graph highlights the overall trend in readability between the two, showcasing the ChatGPT summaries tended to be more readable (with higher FRE scores) than the original abstracts.

The evaluation of ChatGPT-generated abstracts across various study types revealed generally favorable outcomes, with some variation between study categories. Systematic reviews and meta-analyses consistently



Graph. 4. Comparison of Evaluation Metrics for ChatGPT-Generated Abstracts Across Study Types. Scores for each evaluation parameter ranged from 0 to 10, where 0 indicated the lowest or most negative result, and 10 indicated the highest or most positive result

received the highest average ratings across most evaluation metrics, particularly for content accuracy (mean = 8.67), clarity (mean = 8.71), and recommendation to colleagues (mean = 8.78). All papers combined scored between approximately 6.01 and 8.08 across the metrics. Clinical trials showed the best scores in hallucination presence (mean = 6.18) and Systematic reviews and meta-analyses showed the best score in technical term usage (mean = 7.25). Case reports had the lowest scores in several areas, most notably in hallucination presence (mean = 5.10) and technical term usage (mean = 6.5), indicating potential weaknesses in ChatGPT's ability to represent case-specific nuances accurately. However, all study types demonstrated relatively high clarity and satisfaction scores, suggesting that abstracts generated by ChatGPT were generally perceived as readable and useful, (Graph 4).

Discussion

In this study, we assessed ChatGPT's ability to simplify scientific abstracts for public understanding and clinical use. The findings indicated that ChatGPT successfully made the abstracts more readable and accessible, which could potentially save time and help both clinicians and the public better understand complex scientific content. This would allow non-experts to engage with evidence-based health papers, instead of relying on potentially misleading media sources. However, the study also identified significant limitations in ChatGPT's performance, including hallucinations—where the AI added data not found in the original abstracts—and the overuse of technical terminology.

If more studies were conducted on the use of ChatGPT in healthcare, its application could become more effective and beneficial for both clinicians and patients. These studies would provide valuable insights into best practices, allowing for improved integration of ChatGPT into clinical settings. By addressing challenges such as hallucination, customization, privacy, and data security, ChatGPT could become a more reliable tool for managing patient care, ultimately enhancing communication, decision-making, and overall healthcare

outcomes. With continued research, the use of ChatGPT could evolve to better meet the needs of both healthcare providers and patients[50].

To enhance the objectivity and reliability of the evaluation process, we implemented a dual evaluation system, where each paper was reviewed independently by two different evaluators. This approach ensured a more consistent and unbiased assessment. Each paper was assigned to two reviewers from different backgrounds, with the goal of providing a comprehensive evaluation. By having multiple evaluators, we minimized individual biases and strengthened the overall validity of the assessment. Furthermore, inter-rater reliability was assessed by comparing the evaluations of different reviewers for each paper, which reinforced the accuracy and consistency of the results.

From a practical standpoint, the integration of ChatGPT into clinical practice must be approached with careful ethical consideration. While the tool demonstrates strong potential to enhance health communication by simplifying complex scientific content, its use should be guided by clear standards. AI-generated summaries should always be disclosed as such, and their content must be reviewed and validated by qualified healthcare professionals before being shared with patients or the public. To support ethical and effective use, institutions should consider developing guidelines for clinicians on how to evaluate, modify, and use AI outputs appropriately. In addition, educating both healthcare providers and patients about the limitations of AI, including its susceptibility to hallucinations and occasional misuse of technical terms, will be critical to building trust. Establishing these practices will help maximize the benefits of AI in clinical settings while minimizing potential harm, ensuring that tools like ChatGPT serve as an aid—rather than a substitute—for professional expertise.

Advantages of ChatGPT

ChatGPT improves readability and comprehension, providing greater clarity for both clinicians and the public. This saves time and reduces misinterpretation for clinicians while improving the public's understanding of scientific health literature. It allows laypeople to access evidence-based information about their conditions rather than relying on unverified sources such as social media or magazines written by non-healthcare professionals. It enables the public to engage with scientific publications. This can also help media companies and communication departments within healthcare organizations to create content directed towards public consumption. Flesch reading scores were better for the AI summary for the majority of articles (31/43) and showed how authors and publishing companies could use AI in the future to improve content readability.

Disadvantages of ChatGPT

Hallucination and technical terms usage scores in the ChatGPT summaries received scores below the acceptable threshold of 7 out of 10. Hallucinations and the use of technical terminology are considerable weaknesses when considering the real-world use of this tool in healthcare.The score for hallucination presence is 6.01 ± 2.07 for all papers. Clinical Trials were rated 6.18 ± 2.19, Systematic Reviews and Meta-analyses 6.03 ± 2.22, and Case Reports 5.10 ± 0.54. These ratings suggest that, while hallucinations are present across all types of papers, the Case Reports category shows a lower assessment of hallucination presence compared to the other types of papers. At times, ChatGPT shows that these aspects are not strong, that even when the layman abstract is wanted from ChatGPT, it uses technical terms and adds its own interpretation to the information in the main abstract. It was difficult to predict when hallucinations would occur or what elements of the abstract might trigger them.

For example in Shorey et al.'s (2021)[46] study, the response generated by ChatGPT includes additional data that was not mentioned in the abstract, which could be classified as "hallucination". While the goal is for ChatGPT to only generate information that is explicitly stated in the abstract, it appears that ChatGPT extrapolated further details from the full text, even though it was only expected to summarize the abstract. This misalignment between the abstract and the model's output highlights a limitation in the model's ability to strictly adhere to the available information in a concise form. This type of hallucination could lead to misinformation if users assume that the additional details provided by the model are part of the original abstract or are accurate without verification. Therefore, it's crucial to validate the generated text against the actual sources to ensure its accuracy and prevent the spread of incorrect information.

In reviewing the study "How active are stroke patients in physiotherapy sessions and is this associated with stroke severity?" by James[5]we identified several instances where the ChatGPT-generated lay summary introduced hallucinations, simplifications, or overgeneralizations compared to the original abstract. The ChatGPT-generated lay summary successfully conveys the core findings of the study but introduces several notable differences from the original abstract. First, while the summary accurately states that patients with more severe strokes engaged in less exercise during physiotherapy, it simplifies the original finding that this difference relates specifically to the *percentage* of session time spent in active exercise—this nuance is lost in the lay phrasing "spent less time moving." Additionally, the summary states that physiotherapy sessions were shorter than planned "by several minutes," a detail not mentioned in the original and therefore a minor hallucination. A more significant inaccuracy appears in the explanation that patients with severe strokes "spent more of the session being passive (e.g., resting)," which is not supported by data in the abstract and represents an inferred or hallucinated interpretation. The summary also generalizes the study's implications -for example, advising therapists to adapt their methods to help patients "get the most out of their sessions." This phrasing softens the original evidence-based recommendation that therapists should intentionally modify their approach to maximize activity, particularly for those with severe strokes. Furthermore, the summary fails to mention that the study was conducted in a single UK acute stroke unit, omitting an important contextual limitation. Overall, while the lay summary is accessible and broadly accurate, it simplifies certain concepts, introduces a few inferred details not present in the source, and overlooks contextual boundaries that are critical for interpreting the findings appropriately.

It is crucial to thoroughly explain the concept of hallucination, as it represents the most challenging aspect of ChatGPT's performance in our study. Specifically, ChatGPT occasionally generated information that was not

present in the original abstract. While the added content was related to the main abstract, it reflected the model's own interpretations rather than a direct reflection of the source material. This highlights the dual nature of ChatGPT: it is not merely an AI that copies and pastes information but rather one that analyzes, interprets, and draws conclusions. While such capabilities can be advantageous, they can also lead to unintended or inaccurate outcomes.

Ahmad et al. (2023)[51] conducted a study to explore ChatGPT's ability to differentiate correct from incorrect information and its effectiveness in teaching language and literature to undergraduate English students. Using qualitative methods, the research found that ChatGPT's inconsistent responses to contextual questions make it an unreliable tool for language learning. In another study that evaluates the performance of GPT-3.5 and GPT-4 versions of ChatGPT in discussing economic concepts using prompts generated from topics in the Journal of Economic Literature. While ChatGPT shows strong capabilities in providing general summaries, it frequently cites non-existent references. Over 30% of the citations from GPT-3.5 are fabricated, with only a slight improvement in GPT-4. Moreover, the accuracy of the model declines as prompts become more detailed. Quantitative evidence of errors in ChatGPT's output underscores the need for verifying content generated by large language models[52]. Chelli et al. (2024)[53] conducted study to assess the performance of large language models (LLMs) like ChatGPT and Bard in conducting systematic reviews by comparing their ability to replicate human-conducted reviews in the field of shoulder rotator cuff pathology. The study found that LLMs had low precision and high hallucination rates, with ChatGPT showing a hallucination rate of 39.6% for GPT-3.5 and 28.6% for GPT-4, and Bard having an even higher rate of 91.4%. The results suggest that while LLMs can assist in literature searches, their high error rates and hallucinations make them unreliable for conducting systematic reviews without thorough validation by researchers[53]. Communication departments and media agencies still need to review the entire paper to ensure that the AI summary is accurate. Authors of these publications can themselves generate AI summaries once these articles are published. Publication houses can also create an AI generated summary to accompany the research articles to be published for general audiences.

With the exception of hallucinations and technical jargon, systematic reviews produced higher ratings across the evaluated parameters than other types of studies. As this work is already a synthesis and summary of other research works, it may be easier to summarize the abstracts of this type of research. It may also be possible that some of the hallucinations observed in the abstracts were part of the original studies included in the systematic review but not overtly referenced in the systematic review abstract.

The usage of technical terms in ChatGPT-generated summaries was observed to be higher than the acceptable rate, which may lead to difficulties in understanding the abstract for readers, particularly those without specialized knowledge. While this issue is less critical compared to hallucinations, where fabricated or inaccurate information is presented-it still poses a significant barrier to accessibility. Overly technical language may alienate non-expert readers, undermining the goal of simplifying complex scientific content for broader audiences. In future iterations of ChatGPT, reducing the use of excessive technical jargon will be crucial to improve its utility. Striking a balance between accuracy and simplicity can make AI-generated summaries more effective in knowledge dissemination and public engagement.

ChatGPT's performance varies by study type

The evaluation demonstrates that ChatGPT's performance in generating scientific abstracts is variable, ranging from moderate to good depending on the study type. Notably, systematic reviews and meta-analyses achieved the highest average ratings across most evaluation metrics, including content accuracy (mean = 8.67), clarity (mean = 8.71), and recommendation to colleagues (mean = 8.78). These results suggest that the structured and comprehensive nature of systematic reviews may align well with ChatGPT's strengths in synthesizing and presenting information. In contrast, while clinical trials received relatively moderate ratings overall, they performed best in minimizing hallucinations (mean = 6.18), indicating a potentially more fact-bound abstract generation for this study type. Case reports, on the other hand, scored lowest in hallucination presence (mean = 5.10) and technical term usage (mean = 6.5), likely due to their unique, individualized content that may challenge ChatGPT's generalization abilities. Despite these variations, all study types received favorable clarity and satisfaction scores, supporting ChatGPT's usefulness in generating readable abstracts. These findings suggest that while ChatGPT shows promise in supporting scientific writing, especially for systematic reviews and meta-analyses, caution should be taken with case-specific content, and further refinement or human oversight remains important.

The variation in ChatGPT's performance across study types appears to be influenced by differences in abstract complexity and structure. Systematic reviews and meta-analyses, which tend to follow standardized formats and present synthesized findings, were associated with higher clarity and accuracy scores. In contrast, case reports—often more narrative and context-specific—presented greater challenges, likely due to their unique, less predictable content. These findings highlight the importance of understanding how input characteristics affect AI-generated summaries. Moreover, the frequent use of technical language in outputs underscores the need for improved prompt strategies or model fine-tuning aimed at reducing jargon. Future work should explore targeted interventions such as domain-specific prompt engineering or real-time readability feedback to enhance accessibility for non-expert readers, ultimately improving the practical utility of ChatGPT in clinical and public-facing communication.

Reducing Hallucinations and Technical Jargon:

To improve the reliability of ChatGPT-generated content in clinical and research settings, several strategies can be implemented. First, post-generation validation by subject-matter experts remains essential. Clinicians and researchers should cross-check the AI-generated summaries with the original source to ensure accuracy and consistency. Second, using domain-specific fine-tuning or prompting techniques may help reduce hallucinations.

When ChatGPT is provided with a more structured prompt—such as explicitly instructing it to "only include information from the abstract and not infer or add details"—it tends to stay more grounded.

In terms of technical jargon, custom prompts tailored to the target audience can help simplify language. For example, asking ChatGPT to "rewrite the summary in plain language suitable for patients or the general public" often improves accessibility. Future research could test the effectiveness of these strategies through controlled trials or develop hybrid systems that combine AI summaries with real-time validation.

Two-Phase validation framework for reducing hallucinations in ChatGPT-Generated summaries

To ensure the accuracy and reliability of ChatGPT-generated abstracts in clinical and academic contexts, we propose a structured two-phase validation protocol: Pre-Generation Steps (Before Using ChatGPT) and Post-Generation Validation (After Getting the AI Output). This protocol includes both expert oversight and an internal double-check mechanism using ChatGPT itself.

Phase 1: Before Using ChatGPT (Pre-Generation Steps).

1. **Prepare a Structured Prompt**:

- Clearly instruct ChatGPT to *only use information from the original abstract* (e.g., *"Summarize the abstract strictly without adding interpretations, assumptions, or inferred details"*).
- Adjust the tone and language depending on the target audience (e.g., *"Use plain language suitable for patients"* or *"Use clinical terms suitable for specialists"*).

2. **Restrict Input to Abstract Only**:

- Ensure that only the original abstract text is provided. Avoid providing additional data that may cause the model to generalize or hallucinate.

Phase 2: After Using ChatGPT (Post-Generation Validation and Oversight).

1. **Double-Check with ChatGPT Itself (AI Self-Validation)**:

- Ask ChatGPT: *"Compare the generated summary with the original abstract. Identify if there is any information in the summary that is not explicitly present in the original abstract."* This step can help automatically detect hallucinated content, especially subtle additions or inferred ideas.

2. **Expert Cross-Validation**:

- Two independent reviewers (ideally a clinician and a researcher) should: Compare the original abstract and the ChatGPT summary line-by-line. Use a standardized checklist to identify: Factual discrepancies, inferred conclusions, fabricated numerical values or settings, misuse of technical terminology.

3. **Flag and Revise Hallucinations**:

- Any statement not traceable to the original abstract should be:

– Deleted or rephrased,
– Or clearly labeled as inferred (if applicable for internal use, not public dissemination).

4. **Final Approval for Use**:

- Only summaries that pass expert validation with ≥90% factual alignment should be used for patient education, public communication, or research dissemination.

A cross-sectional study evaluated the effectiveness of large language models (LLMs)-including ChatGPT, Claude, Copilot, and Gemini-in generating plain language summaries (PLSs) and simplifying medical content. Thirty human-written PLSs were compared with chatbot-generated counterparts. Readability was assessed using the Flesch Reading Ease score, and understandability via the Flesch-Kincaid grade level. Additionally, three authors independently rated the summaries using a predefined seven-item quality rubric. The results showed that, compared to human-written PLSs, LLMs produced summaries with significantly lower Flesch-Kincaid grade levels ($p < 0.0001$), indicating greater accessibility. Except for Copilot, all chatbots also achieved higher Flesch Reading Ease scores. Quality ratings were comparable across groups—for example, ChatGPT scored 8.8 ± 0.34, closely matching human-written PLSs (8.89 ± 0.26). Similarly, another study evaluating ChatGPT-3.5's performance in simplifying radiological reports found high accuracy in detailing (94.17%) and effective removal of technical jargon, though limitations were noted in patient-directed recommendations and translations. They reported that the current free version of ChatGPT-3.5 was able to simplify radiological reports effectively, removing technical jargon while preserving essential diagnostic information[54,55]. These findings support the potential of LLMs in enhancing communication of scientific and clinical information while reinforcing the need for human oversight to ensure factual correctness.

Limitations of ChatGPT

ChatGPT provides results by taking into account the individual's previous searches. When the different reviewers generated the second abstract summary, the content of the results were found to be consistent although the style of writing and order of the study results was often different.

Conclusion

ChatGPT has significant potential for the dissemination of evidence and knowledge translation in the clinical setting. Our study demonstrated enhanced readability scores and an acceptable degree of accuracy when comparing the original abstracts to the AI summary. However, caution should be exercised, and education provided to colleagues and patients, regarding the risk of hallucination and incorporation of technical jargon which may make the results challenging to interpret. Future studies are warranted to establish the reliability, validity and safety of ChatGPT and other AI tools as a real-world tool for knowledge translation.

Data availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Received: 14 March 2025; Accepted: 8 July 2025
Published online: 29 September 2025

References

1. OpenAI & ChatGPT Optimizing language models for dialogue. OpenAI. (2023). Available from: https://openai.com/chatgpt
2. Dave, T., Athaluri, S. A. & Singh, S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front. Artif. Intell.* **6**, 1169595. https://doi.org/10.3389/frai.2023.1169595 (2023).
3. Jiang, F. et al. Artificial intelligence in healthcare: past, present, and future. *Stroke Vascular Neurol.* **2** (4), 230–243. https://doi.org/10.1136/svn-2017-000101 (2017).
4. Islam, M. R., Urmi, T. J., Mosharrafa, R. A., Rahman, M. S. & Kadir, M. F. Role of ChatGPT in health science and research: A correspondence addressing potential application. *Health Sci. Rep.* **6** (10), e1625. https://doi.org/10.1002/hsr2.1625 (2023).
5. James, J. & McGlinchey, M. P. How active are stroke patients in physiotherapy sessions and is this associated with stroke severity? *Disabil. Rehabil.* **44** (16), 4408–4414. https://doi.org/10.1080/09638288.2021.1907459 (2022).
6. Rahayu, U. B., Wibowo, S., Setyopranoto, I. & Hibatullah Romli, M. Effectiveness of physiotherapy interventions in brain plasticity, balance and functional ability in stroke survivors: A randomized controlled trial. *NeuroRehabilitation* **47** (4), 463–470. https://doi.org/10.3233/NRE-203210 (2020).
7. Kanase Suraj, B., Varadharajulu, G., Salunkhe Pragati, V. & Burungale Mayuri, D. Role of physiotherapy on quality of life in stroke Survivors – A systematic review. *Indian J. Forensic Med. Toxicol.* **14** (2), 226–230 (2020).
8. Gamble, K., Chiu, A. & Peiris, C. Core stability exercises in addition to usual care physiotherapy improve stability and balance after stroke: A systematic review and meta-analysis. *Arch. Phys. Med. Rehabil.* **102** (4), 762–775. https://doi.org/10.1016/j.apmr.2020.09.388 (2021).
9. Veldema, J. & Jansen, P. Resistance training in stroke rehabilitation: systematic review and meta-analysis. *Clin. Rehabil.* **34** (9), 1173–1197. https://doi.org/10.1177/0269215520932964 (2020).
10. Navarro-López, V., Molina-Rueda, F., Jiménez-Jiménez, S., Alguacil-Diego, I. M. & Carratalá-Tejada, M. Effects of transcranial direct current stimulation combined with physiotherapy on gait pattern, balance, and functionality in stroke patients. A systematic review. *Diagnostics* **11** (4), 656. https://doi.org/10.3390/diagnostics11040656 (2021).
11. Smedes, F. & da Silva, L. G. Motor learning with the PNF-concept, an alternative to constrained induced movement therapy in a patient after a stroke: A case report. *J. Bodyw. Mov. Ther.* **23** (3), 622–627. https://doi.org/10.1016/j.jbmt.2018.05.003 (2019).
12. Miclaus, R. et al. Non-immersive virtual reality for post-stroke upper extremity rehabilitation: A small cohort randomized trial. *Brain Sci.* **10** (9), 655. https://doi.org/10.3390/brainsci10090655 (2020).
13. Nave, A. H. et al. Physical fitness training in patients with subacute stroke (PHYS-STROKE): multicentre, randomized controlled, endpoint-blinded trial. *BMJ* **366**, l5101. https://doi.org/10.1136/bmj.l5101 (2019).
14. Dagal, R. & Thimoty, R. Effect of physiotherapy on hand rehabilitation in acute ischemic stroke survivor: A case report. *J. Pharm. Res. Int.* **34** (1A), 28–32 (2022).
15. Aebischer, B., Elsig, S. & Taeymans, J. Effectiveness of physical and occupational therapy on pain, function and quality of life in patients with trapeziometacarpal osteoarthritis–A systematic review and meta-analysis. *Hand Therapy.* **21** (1), 5–15 (2016).
16. Ahn, S. N. Effectiveness of occupation-based interventions on performance's quality for hemiparetic stroke in community-dwelling: A randomized clinical trial study. *NeuroRehabilitation* **44** (2), 275–282 (2019).
17. Maitra, K. et al. Five-year retrospective study of inpatient occupational therapy outcomes for patients with multiple sclerosis. *Am. J. Occup. Therapy.* **64** (5), 689–694 (2010).
18. Ritter, V. C. & Bonsaksen, T. Improvement in quality of life following a multidisciplinary rehabilitation program for patients with parkinson's disease. *J. Multidisciplinary Healthc.* **20**, 219–227 (2019).
19. Cole, T. et al. Outcomes after occupational therapy intervention for traumatic brachial plexus injury: A prospective longitudinal cohort study. *J. Hand Ther.* **33** (4), 528–539 (2020).
20. Ceylan, İ. et al. The effectiveness of mobilization with movement on patients with mild and moderate carpal tunnel syndrome: A single-blinded, randomized controlled study. *J. Hand Ther.* **36** (4), 773–785 (2023).
21. Werner, F. W. et al. Scaphoid tuberosity excursion is minimized during a dart-throwing motion: A Biomechanical study. *J. Hand Ther.* **29** (2), 175–182 (2016).
22. Kim, J. K., Al-Dhafer, B., Shin, Y. H. & Joo, H. S. Effect of pre-treatment expectations on post-treatment expectation fulfillment or outcomes in patients with distal radius fracture. *J. Hand Ther.* **36** (1), 97–102 (2023).
23. Schwartz, D. A. & Schofield, K. A. Utilization of 3D printed orthoses for musculoskeletal conditions of the upper extremity: A systematic review. *J. Hand Ther.* **36** (1), 166–178 (2023).
24. Cochrane, S. K., Calfee, R. P., Stonner, M. M. & Dale, A. M. The relationship between depression, anxiety, and pain interference with therapy referral and utilization among patients with hand conditions. *J. Hand Ther.* **35** (1), 24–31 (2022).
25. Hakverdioğlu Yönt, G., Kisa, S. & Princeton, D. M. Physical restraint use in nursing Homes-Regional variances and ethical considerations: A scoping review of empirical studies. *Healthcare* **11** (15), 2204. https://doi.org/10.3390/healthcare11152204 (2023).
26. Mehta, G. et al. Impact of diabetes on inpatient length of stay in adult mental health services in a community hospital setting: A retrospective cohort study. *Can. J. Diabetes.* **46** (7), 678–682 (2022).

27. Sa, Z. et al. Impact of mental disorders on unplanned readmissions for congestive heart failure patients: a population-level study. *ESC Heart Fail.* **11** (2), 962–973. https://doi.org/10.1002/ehf2.14644 (2024).
28. Lapierre, N. et al. Exergame-assisted rehabilitation for preventing falls in older adults at risk: A systematic review and meta-analysis. *Gerontology* **69** (6), 757–767 (2023).
29. Hawley-Hague, H. et al. Using smartphone technology to support an effective home exercise intervention to prevent falls amongst community-dwelling older adults: the TOGETHER feasibility RCT. *Gerontology* **69** (6), 783–798 (2023).
30. Rafferty, M. et al. Promoting Evidence-Based practice: the influence of novel structural change to accelerate translational rehabilitation. *Arch. Phys. Med. Rehabil.* **104** (8), 1289–1299. https://doi.org/10.1016/j.apmr.2023.02.014 (2023).
31. Yang, Y. et al. Construction and application of a nursing human resource allocation model based on the case mix index. *BMC Nurs.* **22** (1), 466 (2023).
32. Jallad, S. T. & Işık, B. Transitioning nursing students' education from traditional classroom to online education during the COVID-19 pandemic: A case study applied to the Meleis trial. *Florence Nightingale J. Nurs.* **29** (1), 124 (2021).
33. Or, C. K. et al. Improving self-care in patients with coexisting type 2 diabetes and hypertension by technological surrogate nursing: randomized controlled trial. *J. Med. Internet. Res.* **22** (8), e22518 (2020).
34. Tuomikoski, A-M. et al. Nurses' experiences of their competence at mentoring nursing students during clinical practice: A systematic review of qualitative studies. *Nurse Educ. Today*. **85**, 104258 (2020).
35. Kringle, E. A. et al. Development and feasibility of a sedentary behavior intervention for stroke: A case series. *Top. Stroke Rehabil.* **26** (6), 456–463. https://doi.org/10.1080/10749357.2019.1623437 (2019).
36. Poltawski, L. et al. Informing the design of a randomised controlled trial of an exercise-based programme for long-term stroke survivors: lessons from a before-and-after case series study. *BMC Res. Notes*. **6**, 324. https://doi.org/10.1186/1756-0500-6-324 (2013).
37. Sarikaya, H., Ferro, J. & Arnold, M. Stroke prevention - Medical and lifestyle measures. *Eur. Neurol.* **73** (3–4), 150–157. https://doi.org/10.1159/000367652 (2015).
38. Lennon, O., Galvin, R., Smith, K., Doody, C. & Blake, C. Lifestyle interventions for secondary disease prevention in stroke and transient ischemic attack: A systematic review. *Eur. J. Prev. Cardiol.* **21** (8), 1026–1039. https://doi.org/10.1177/2047487313481756 (2014).
39. Lund, A., Michelet, M., Sandvik, L., Wyller, T. & Sveen, U. A lifestyle intervention as a supplement to a physical activity programme in rehabilitation after stroke: A randomized controlled trial. *Clin. Rehabil.* **26** (6), 502–512. https://doi.org/10.1177/0269215511429473 (2012).
40. Kono, Y. et al. Secondary prevention of new vascular events with lifestyle intervention in patients with noncardioembolic mild ischemic stroke: A single-center randomized controlled trial. *Cerebrovasc. Dis.* **36** (2), 88–97. https://doi.org/10.1159/000352052 (2013).
41. Matz, K. et al. Multidomain lifestyle interventions for the prevention of cognitive decline after ischemic stroke: randomized trial. *Stroke* **46** (10), 2874–2880. https://doi.org/10.1161/STROKEAHA.115.009992 (2015).
42. Olaiya, M. T. et al. Community-based intervention to improve cardiometabolic targets in patients with stroke: A randomized controlled trial. *Stroke* **48** (9), 2504–2510. https://doi.org/10.1161/STROKEAHA.117.017499 (2017).
43. Deijle, I. A., Van Schaik, S. M., Van Wegen, E. E., Weinstein, H. C. & Van den Kwakkel, G. Lifestyle interventions to prevent cardiovascular events after stroke and transient ischemic attack: systematic review and meta-analysis. *Stroke* **48** (1), 174–179. https://doi.org/10.1161/STROKEAHA.116.013794 (2017).
44. Altobelli, E., Angeletti, P. M., Rapacchietta, L. & Petrocelli, R. Overview of meta-analyses: the impact of dietary lifestyle on stroke risk. *Int. J. Environ. Res. Public Health*. **16** (19), 3582. https://doi.org/10.3390/ijerph16193582 (2019).
45. Eslamian, J., Moeini, M. & Soleimani, M. Challenges in nursing continuing education: A qualitative study. *Iran. J. Nurs. Midwifery Res.* **20** (3), 378 (2015).
46. Shorey, S. & Wong, P. Z. E. A qualitative systematic review on nurses' experiences of workplace bullying and implications for nursing practice. *J. Adv. Nurs.* **77** (11), 4306–4320 (2021).
47. Dos Santos, L. M. Stress, burnout, and low self-efficacy of nursing professionals: A qualitative inquiry. *Healthc. (Basel)*. **8** (4), e105. https://doi.org/10.3390/healthcare8040105 (2020).
48. Fajardo, M. A., Weir, K. R., Bonner, C., Gnjidic, D. & Jansen, J. Availability and readability of patient education materials for deprescribing: an environmental scan. *Br. J. Clin. Pharmacol.* **85** (7), 1396–1406. https://doi.org/10.1111/bcp.13912 (2019).
49. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **32** (3), 221–233. https://doi.org/10.1037/h0057532 (1948).
50. Miao, J., Thongprayoon, C., Fülöp, T. & Cheungpasitporn, W. Enhancing clinical decision-making: optimizing chatgpt's performance in hypertension care. *J. Clin. Hypertens.* **26** (5), 588 (2024).
51. Ahmad, Z., Kaiser, W. & Rahim, S. Hallucinations in chatgpt: an unreliable tool for learning. *Rupkatha J. Interdisciplinary Stud. Humanit.* **15** (4). https://doi.org/10.21659/rupkatha.v15n4.17 (2023).
52. Buchanan, J., Hill, S. & Shapoval, O. ChatGPT hallucinates non-existent citations: evidence from economics. *Am. Econ.* **69** (1), 80–87. https://doi.org/10.1177/05694345231218454 (2024).
53. Chelli, M. et al. Hallucination rates and reference accuracy of ChatGPT and bard for systematic reviews: comparative analysis. *J. Med. Internet. Res.* 26. https://doi.org/10.2196/53164 (2024).
54. Mondal, H. et al. Assessing the capability of large Language model chatbots in generating plain Language summaries. *Cureus* **17** (3), e80976. https://doi.org/10.7759/cureus.80976 (2025).
55. Sarangi, P. K. et al. Assessing chatgpt's proficiency in simplifying radiological reports for healthcare professionals and patients. *Cureus* **15** (12), e50881. https://doi.org/10.7759/cureus.50881 (2023).

Acknowledgements

Ethical Considerations: Not applicable, as this study does not involve human participants, literature and all studies included in this study are cited in the references.

Author contributions

Authorship contribution statement Esra Dogru-Huzmeli: Supervision. Esra Dogru-Huzmeli, Sarah Moore-Vasram, Chetan Phadke: Writing – review & editing, Esra Dogru-Huzmeli, Sarah Moore-Vasram, Chetan Phadke, Erfan Shafiee, Amanullah Shabbir: Writing – original draft, Methodology.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to E.D.-H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.