Journal of Clinical Epidemiology 196 (2026) 112309

ORIGINAL RESEARCH

Automated approaches to identifying clinical trials based on title and abstract in the field of physiotherapy: a comparative analysis

Wing S. Kwok[a,b,*], Geraldine Wallbank[a,b], Philip Hodgson[c,d], Thomas Schrader[e], Lexuan Shao[b], Mark Elkins[f], Junior Fandim[g], Julia Scott[f], Catherine Sherrington[a,b], Adrian C. Traeger[a,b]

[a]*Institute for Musculoskeletal Health, Sydney Local Health District, Sydney, Australia*
[b]*School of Public Health, Faculty of Medicine and Health, The University of Sydney, Australia*
[c]*Physiotherapy Department, Tees, Esk and Wear Valleys NHS Foundation Trust, West Park Hospital, Edward Pease Way, Darlington, United Kingdom, GB-DL2 2TS*
[d]*School of Science, Technology and Health, York St John University, Lord Mayor's Walk, York GB-YO31 7EX*
[e]*Department of Informatics and Media, Brandenburg University of Applied Sciences, Brandenburg, Germany*
[f]*Faculty of Medicine and Health, University of Sydney, Australia*
[g]*Department of Physical Therapy, Universidade Cidade de São Paulo, São Paulo, Brazil*

Accepted 27 April 2026; Published online 30 April 2026

Abstract

Objectives: To compare accuracy, precision, recall, F1, and time spent using commercial tools to identify physiotherapy trials based on title and abstract, compared with a human approach.

Study Design: This study compared two approaches for title and abstract screening of 10,793 newly published records. In the reference standard human approach, two reviewers independently screened records using prespecified rules to assess relevance to physiotherapy. A third person resolved disagreements. We evaluated three large language models (LLMs) (gpt-4o, gpt-4.5, and gpt-4-turbo) within two commercial, web-based tools (ChatGPT and Copilot). Outcomes were accuracy (proportion of records that model correctly identified as relevant or irrelevant), precision (proportion of records identified as relevant that were considered as relevant by human approach), recall (the proportion of all actual relevant records that the model successfully identified), F1 (harmonic mean of precision and recall), and time spent. Exploratory analyses compared the performance of the commercial tools with local approaches, including local LLMs implementation, machine learning, and natural language processing.

Results: Commercial tools showed comparable performance across all metrics (ChatGPT vs Copilot: accuracy: 83% vs 86%; precision: 44% vs 48%; recall: 88% vs 87%; F1: 59% vs 62%). The total time spent using commercial tools with a labeled dataset was equivalent to 37% of the time required for the human-only screening process. Exploratory analysis showed that the Application Programming Interface–based implementation has comparable performance (accuracy: 82%; precision: 42%; recall: 93%; F1: 58%). Yet, LLM-based models demonstrated lower performance compared with other local, custom-adapted automation approaches such as machine learning and natural language processing.

Conclusion: This proof-of-concept study demonstrates that commercial web-based LLMs may have sufficient accuracy to support title and abstract screening and substantially reduce the time to identify field-specific trials. However, alternative approaches, including machine learning or natural language processing, could achieve screening performance similar to or slightly higher than that of commercial tools, yet they require a series of preprocessing steps for implementation.

Keywords: Physiotherapy; Large language model; Rehabilitation; Machine learning; Natural language processing; Screening

1. Introduction

Screening large volumes of records in research databases is a critical step in evidence synthesis. In systematic reviews, this process is often the most time-consuming aspect [1]. For large evidence databases that index all articles relevant to a subject area, the routine screening of search

Registration: OSF https://osf.io/94xza/.

* Corresponding author. Institute for Musculoskeletal Health, University of Sydney and Sydney Local Health District, Level 10N, King George V Building, Royal Prince Alfred Hospital (C39), PO Box M179, Missenden Road, Camperdown, NSW 2050, Australia.
E-mail address: venisa.kwok@sydney.edu.au (W.S. Kwok).

https://doi.org/10.1016/j.jclinepi.2026.112309

What is new?

Key findings

- Commercial web-based large language models could reduce the workload to identify physiotherapy-specific trials based on title and abstract (n = 10,793).

What this adds to what was known?

- Commercial web-based large language models may underperform compared with machine learning and natural language processing models.

What is implication, what should change now?

- Automated approaches, including large language models, machine learning, and natural language processing, can support title and abstract screening.
- Resource requirements, including a series of preprocessing steps, may limit the scalability of machine learning and natural language processing.

results to identify eligible studies is even more burdensome. For example, physiotherapy literature has grown at an exponential rate, and the Physiotherapy Evidence Database (PEDro; www.pedro.org.au), a freely available database, now indexes over 66,000 records [2,3]. These records have been selected by screening the titles and abstracts of a pool of approximately 875,000 records. With the ongoing surge in published evidence, manually identifying clinical trials relevant for inclusion in PEDro has become increasingly challenging.

Technological advances in artificial intelligence (AI) have expanded the range of computational approaches available to support and automate evidence selections. Early application includes traditional machine learning techniques, which apply supervised or unsupervised algorithms to data to learn patterns and make task-specific predictions [4]. Subsequently, transformer-based language models have improved semantic text representation by capturing the meaning of words and sentences within their context, enabling more accurate text classification and information extraction [5,6]. Building on these advances, large language models (LLMs) leverage training on large-scale textual data to understand, extract, and generate human-like text [7], with growing applications in medical text processing (eg, summarizing and translating medical text, answering medical questions) [8] and supporting systematic reviews [9]. Existing commercial, web-based LLMs or 'generative AI' systems offer prompt-based interactions to rapidly evaluate texts and language [10]. This approach is easily accessible, reduces technical complexity without requiring specialist programming skills, and potentially provides a structured, repeatable output. LLMs can also be deployed via Application Programming Interface (API)−based architectures to further improve repeatability.

Growing evidence shows that LLMs can support evidence screening across multiple disciplines, including environmental science [11], clinical medicine (eg, pharmacology [12], neurosurgery [13], sepsis, and septic shock), [14] and mental health [15−17]. For example, a study found that LLM-assisted screening across diverse systematic review topics, including the field of musculoskeletal physiotherapy, reported a workload reduction of at least 33% while accuracy reaching up to 92% [18]. Yet, the study did not specifically examine the use of LLMs to identify relevant trials in the broader field, which, in addition to musculoskeletal, includes cardiopulmonary, neurology, and men's and women's health, among other specialty areas. Such complexity demands further validation of automated approaches for evidence screening. Key metrics to determine whether an LLM is useful include accuracy (ie, proportion of records correctly classified), precision (ie, the proportion of records the model identified as relevant that are truly relevant), and recall (ie, the proportion of all actually relevant records that the model successfully identified). We compared the accuracy, precision, recall, F1, and time spent of commercial, web-based LLMs for identifying physiotherapy clinical trials with the traditional human approach.

2. Method

This study evaluated the performance of readily available, commercial, web-based LLMs (ChatGPT and Copilot) in screening the titles and abstracts of clinical trial articles in the field of physiotherapy for eligibility of inclusion in the PEDro. The results generated by these commercial web-based LLMs were compared with a human approach, which was considered the reference standard. The protocol for this study was recorded on the Open Science Framework (https://osf.io/94xza/) before any analysis. The study is reported according to the generative AI tools in medical research checklist [19].

2.1. Data source

A standard targeted search of biomedical databases, conducted monthly between August and November 2024, was used to identify records relevant for indexing in PEDro. The PEDro team searches the following databases: Medline via Ovid, American Psychological Association PsycINFO via Ovid, AMED via Ovid, Embase via Ovid, Cumulative Index of Nursing and Allied Health Literature via EBSCO-host, and Cochrane Central Register of Controlled Trials. This study included records that were written in English.

2.2. Procedures

We used two different approaches to screening the same set of articles: 1) a human approach (reference standard) and 2) a commercial web-based LLM-based approach. We defined relevance to physiotherapy according to the PEDro inclusion criteria, as stated on the PEDro website [3]. In brief, a trial was considered relevant if a trial:

- Compared at least two interventions and reported the comparative effectiveness, with at least one intervention that is currently or could become a part of physiotherapy practice;
- Was conducted in participants who were representative or were intended to be representative in the course of physiotherapy practice;
- Used random or intended-to-be-random allocation to interventions.

2.3. Human approach

A pair of trained PEDro staff members/affiliates (authors P.H., J.F., J.S.) independently screened the titles and abstracts to determine whether each trial record met the inclusion criteria. A third reviewer (W.K. or G.W.) resolved any disagreements. This approach is commonly used in systematic reviews, including Cochrane Reviews, and was considered as the reference standard [20,21]. The time each reviewer spent screening was recorded.

2.4. Commercial, web-based LLM approach

We used a structured English prompt (Appendix 1). Instructions and questions were added to each commercial tool in a new conversation one by one. We used three LLMs (gpt-4o, gpt-4.5, and gpt-4-turbo) within two commercial web-based (ie, Microsoft Copilot with institutional subscription [22,23] and ChatGPT with and without a subscription [24]). We provided previously screened articles, which were screened using the same approach as the human approach described earlier, from various dataset sizes (n = 100, 1000, 2000, 3000, and 4000) for in-context learning. We assessed whether varying dataset sizes of labeled records affected the model's results. We started with a smaller number of labeled records. The dataset of 4000 labeled records, with 40% of records labeled as included, was used as the primary analysis. The commercial web-based LLMs provided the predicted labels for each record. To ensure robustness of results and minimize the risk of prior interactions, we started a new chat for each analysis and disabled memory features (where applicable). We recorded the total time required, including dataset preparation and testing. We reran the analysis 2 months later to evaluate whether the approach maintained reproducibility over time.

2.5. Exploratory analysis

We conducted further exploratory analyses to compare other existing approaches to automation of screening, and this included 1) an API-based implementation with retrieval-augmented generation (RAG), 2) a machine learning approach, and 3) a natural language processing. All model outputs were also evaluated against the consensus labels from the human approach. Details of the codes and setup of these approaches are publicly available at https://github.com/Kenn0918/SydneyUniPEDro.

2.6. API-based implementation with RAG

An API-based implementation was developed to replicate the initial study design using programmatic access to LLM. Specifically, gpt-4o and gpt-4-turbo were accessed via the OpenAI API under a controlled setting. We did not evaluate gpt-4.5, as it was not available at the time of testing. To approximate the use of the contextual information, an RAG pipeline was implemented [25]. For each input record (title and abstract), the query was embedded and used to retrieve relevant examples from a labeled dataset using similarity-based search within a vector index. A top-k retrieval strategy was applied to select the relevant records, which were then incorporated into the model prompt. The same prompt structure used in the primary analysis was retained to ensure comparability.

2.7. Machine learning approach

We examined two widely used machine learning methods, including 1) Support vector machine (SVM) and 2) logistic regression [26]. Text data (ie, title and abstract) were preprocessed, including a cleansing procedure with lemmatization (transforming a word to its dictionary entity [lemma]) and stop word removal. To address potential class imbalance, we used both balanced/unbalanced training settings for the model development. SVM is a supervised machine learning and was used to find the optimal boundary (a line or hyperplane) that best separates data into different classes by maximizing the margin between the closest points of each class [27]. Logistic regression was used to estimate the probability that a given input belongs to a particular category, using a logistic function to output values between 0 and one that can be interpreted as class membership likelihood [28].

2.8. Natural language processing model

We used Bidirectional Encoder Representations from Transformers (BERT), a model that understands the meaning of words by analyzing their full context in both directions, simultaneously considering words that precede and follow them. Concerning the bidirectional approach of embedded words, BERT is considered effective for text classification tasks [29]. Nevertheless, the classification

results depend on the language model used. We used the ClincalBERT model, a special text cleaning process is not required to apply BERT [30].

2.9. Deviation from the protocol

We planned to use another tool, Document Classification and Topic Extraction Resource [31,32], an online supervised machine learning platform to predict whether a trial record is relevant or not. Unexpectedly, the platform did not allow the predicted file to be downloaded despite multiple attempts. Therefore, we were unable to assess its performance.

2.10. Statistical analysis

The performance of all the computational approaches (ie, commercial web-based LLMs, API-based implementation, machine learning, and natural language processing)

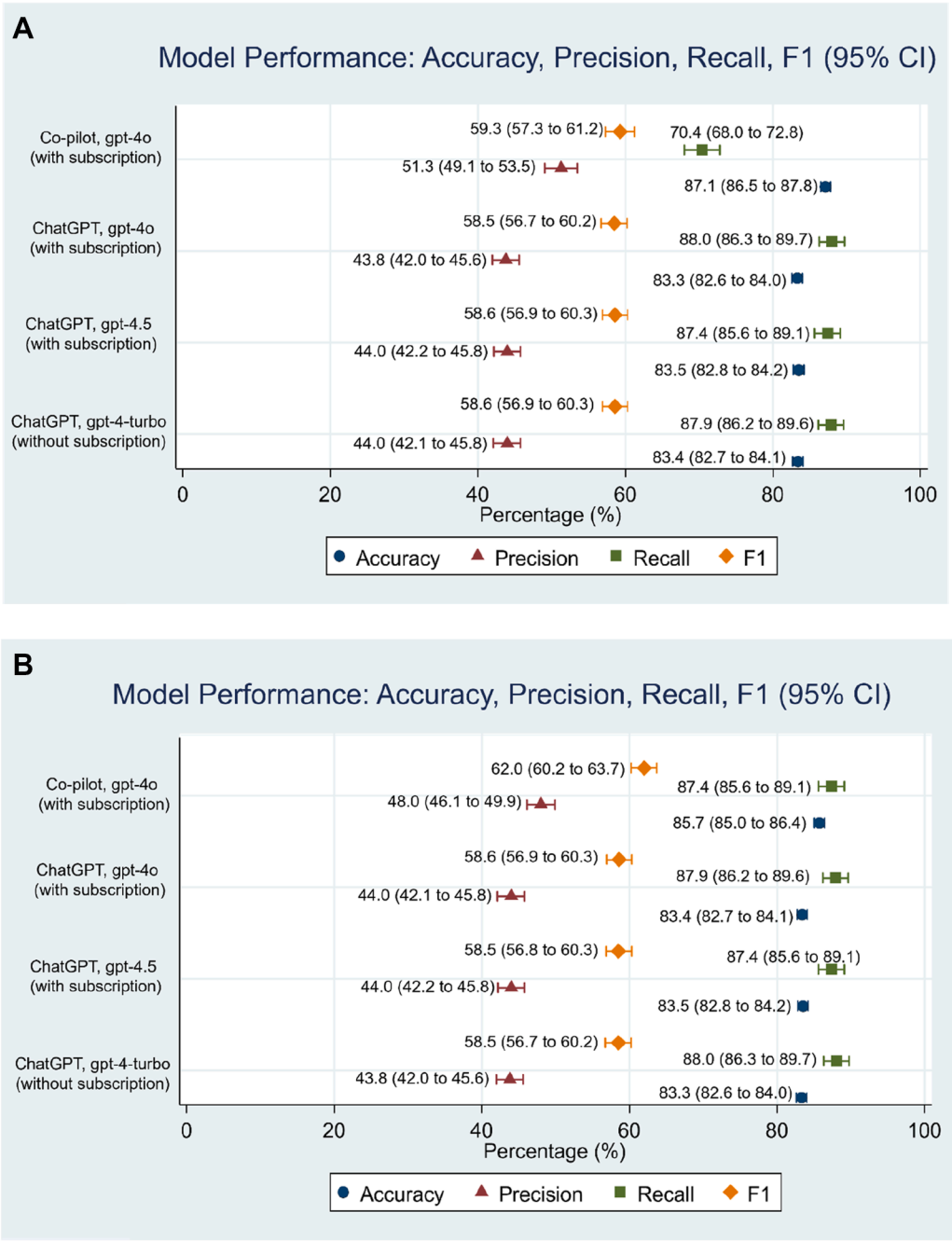

Figure. A, Performance metrics (accuracy, precision, recall, and F1) with 95% CI for ChatGPT and Copilot evaluated on the dataset of 4000 labeled records. B, Performance metrics (accuracy, precision, recall, and F1) with 95% CI for ChatGPT and Copilot evaluated on the dataset of 4000 labeled records in the repeat assessment.

was evaluated against the consensus decision in the human approach, which served as the 'reference standard' for classification metrics. Standard prespecified classification metrics, including accuracy, precision, and recall, were reported as percentages and their 95% confidence interval (CI). To reflect the model's discriminative ability, we have also reported the F1 score (ie, the harmonic mean of precision and recall) [33]. Details of these metric formulae are presented in Appendix 2. We explored agreement between 1) human raters and 2) an experienced staff and the commercial web-based LLM approach using Cohen's kappa. Efficiency was evaluated by comparing the time required for a human vs a commercial web-based LLM approach. Data were compared using STATA 16.0 (StataCorp LLC).

3. Results

A total of 11,766 records were retrieved via targeted searches across databases in the 3-month period. After removal of non-English articles (n = 259) and duplicates using EndNote (n = 714) [34], 10,793 records were screened. Using the human review approach, interrater agreement between reviewers was 89.9% (9703 of 10,793), Cohen's kappa was 0.48 (95% CI 0.46−0.50), indicating moderate agreement [35]. A total of 1439 records (13.3%) were deemed relevant, and 9354 records (86.7%) were deemed irrelevant.

3.1. Primary analysis

Using a dataset of 4000 labeled records, the results of the initial (Fig A) and repeat test (Fig B) using ChatGPT and Copilot were mostly comparable. Overall, the accuracy of ChatGPT and Copilot in the initial test ranged from 83% to 87%. Copilot achieved higher precision (51%, 95% CI 49−53) but lower recall (70%, 95% CI 68−73) than ChatGPT, with similar F1 scores. Different LLMs within ChatGPT showed similar precision and recall. Copilot had a lower recall in the initial test but improved to levels comparable to ChatGPT in the repeat test (Fig B).

With labeled record datasets of varying sizes, Copilot showed very low recall and F1 scores when only 100 (recall 1%; F1 1%) and 1000 (recall 28%; F1 22%) labeled records were used (Appendices 3 and 4). ChatGPT also had its lowest recall with 100 labeled records (63%), compared to 1000 labeled records (86%), after which performance plateaued. On repeat testing, the low accuracy scores and precision of Copilot at 100 and 2000 labeled records were not observed (Appendices 5 and 6). Agreement between experienced staff and the commercial web-based LLM tools ranged from 0.42 to 0.48, indicating moderate agreement [35].

3.2. Time spent

Humans spent 135.5 hours of screening titles and abstracts for 10,793 records (125.5 hours for independent reviewer screening and 10 hours for consensus adjudication). Preparing a dataset of 4000 labeled records using the same approach took 50.2 hours. Uploading the dataset and obtaining responses took approximately 5 minutes with ChatGPT and 7 to 8 minutes with Copilot. Overall, in the analysis of the 4000 labeled dataset, the LLM-based approach required 37% of the time of the human-only screening process.

3.3. Exploratory analysis

Table 1 shows performance metrics comparing the commercial, web-based LLMs with other exploratory approaches to automation. An API-based implementation that retrieved a labeled dataset using similarity-based matching showed the lowest performance in accuracy (70%), precision (30%), and F1 (45%). Incorporating eligibility criteria prompts into the API-based implementation led to consistent improvements across all metrics (accuracy: 82%; precision: 42%; F1: 58%), similar to the primary analysis using ChatGPT and Copilot. Both machine learning and natural language processing approaches had a higher accuracy, precision, and F1, than with the LLM-based approach (Table 1).

4. Discussion

This study evaluated the performance of commercial web-based LLMs (ChatGPT, Copilot) in screening titles and abstracts for clinical relevance. The commercial tools demonstrated some ability to accurately classify records by relevance; however, their overall performance was lower compared with local, custom-adapted automation approaches based on machine learning and natural language processing principles. This suggests that bespoke language models built using specialist skills may perform better at screening tasks than the current, commercially available LLMs.

The accuracy of the commercial web-based LLMs remained above 80% in both initial and repeated testing. Precision was notably lower (<50%) across all these web-based LLM tools, indicating a higher rate of false positives and misclassifying irrelevant records as relevant. A high rate of false positives may be less detrimental to internal validity than false negatives when conducting a systematic review, as excluding relevant studies can lead to selection bias. Irrelevant records (ie, false positives) can be excluded at later stages in a review process (eg, full-text screening stage), whereas missing a relevant record could fundamentally undermine an evidence synthesis. Commercial web-based LLM tools may be useful as a filter, eliminating most irrelevant records and substantially reducing the initial human screening burden, although manual review remains necessary to address false positives. Notably, the false negative rate of the commercial tools was around 1.6% to

Table 1. Performance metrics with 95% CI comparing large language models, machine learning, and natural language processing approaches for title and abstract record selection

	LLMs gpt−4o				Machine learning		NLP
(%)	ChatGPT[a]	Copilot	API-based RAG implementation[b]	API-based RAG implementation with PEDro inclusion criteria[c]	SVM	Logistic regression	BERT-based approach
Accuracy	83.4 (82.7−84.1)	85.7 (85−86.4)	69.8 (68.9−70.7)	81.9 (81.2−82.6)	90.2 (89.7−90.7)	89.1 (88.6−89.5)	92.6 (91.6−93.4)
Precision	44 (42.1−45.8)	48 (46.1−49.9)	29.9 (28.5−31.2)	42 (40.3−43.7)	77.9 (76.8−79.1)	82.3 (81.4−83.1)	79.2 (76.9−81.4)
Recall	87.9 (86.2−89.6)	87.4 (85.6−89.1)	93.9 (92.6−95.1)	93.2 (91.9−94.5)	71.6 (69.8−73.4)	58.4 (55.2−61.5)	85.6 (82.1−88)
F1	58.6 (56.9−60.3)	62 (60.2−63.7)	45.3 (43.8−46.9)	57.9 (56.2−59.7)	74.6 (73.2−76.1)	68.3 (66.1−70.4)	82.3 (79.8−84.3)

API, Application Programming Interface; BERT, Bidirectional Encoder Representations from Transformers; LLMs, large language models; NLP, Natural language processing; RAG, retrieval-augmented generation; SVM, support vector machine.

[a] The performance metrics presented were based on the dataset of 4000 labeled records in the repeat assessment and tested on July 1, 2025 (same results as presented in Appendix 5).

[b] Exploratory analysis: API-based RAG implementation was run on March 21, 2026, overnight to March 22, 2026.

[c] Exploratory analysis: API-based RAG implementation with inclusion criteria was run on March 27, 2026, overnight to March 28, 2026.

1.8% of the total records tested. This figure warrants cautious interpretation because it risks omitting important evidence. Yet, the long-term impact of early omission is mitigated, as missed records in an evidence database such as PEDro are often identified later through community feedback.

Differences in recall and F1 values were observed between the initial and repeat testing with the commercial tools. Copilot demonstrated substantially lower recall and F1 during the initial testing, with significant improvement observed upon repeated testing. This temporal variation underscores an emerging challenge in the application of commercial web-based LLMs and may reflect the fundamentally opaque or ''black-box'' nature of commercial tools. Variability of performance may arise from model updates, deployment changes, or inherent variability in generative outputs. Such differences raise concerns about reproducibility, as identical prompts yielded different results across time points. Recall is particularly critical in systematic evidence screening, where missing potentially relevant records can compromise the validity of a review. Notably, our exploratory analysis using an API-based RAG implementation produced similar results, though it required explicit inclusion criteria prompts. API-based approaches are known to produce more repeatable results.

Several studies have examined approaches to the automation of evidence screening [11−18]. However, to our knowledge, only one study has evaluated LLM-assisted screening within musculoskeletal physiotherapy [18], which represents only a subset of the broader physiotherapy field. This study used gpt-4o with explicit inclusion criteria on how individual inclusion criteria contributed to the final decision of inclusion achieved 82% accuracy, 87% recall, but 16% precision, reflecting a large number of false positives [18]. Our study provides a proof-of-concept for achieving a more balanced performance, maintaining similar accuracy (83%) and recall (88%) with moderate precision (44%). One possible reason for the differences in results is the use of previously screened records for in-context learning, which, in our study, may have enabled the model to infer patterns.

Taken together, the observed accuracy, efficiency, and comparable agreement between human−human approach and human−web-based LLMs observed in this study suggest that commercial, web-based LLMs have the potential to support title and abstract screening to identify relevant records. However, our exploratory analyses demonstrated that alternative approaches, such as machine learning and natural language processing, could achieve higher agreement. Although these other approaches outperformed the commercial tools, they require significant effort to implement. In contrast, the LLM approach is more intuitive and can be used effectively by nonexperts. Future research could explore strategies for using LLMs to perform full-text article screening, as having richer contextual information may improve LLMs performance. It will also be interesting to investigate hybrid screening workflows that combine LLM-based and traditional machine learning or natural language processing techniques as independent parallel reviewers, similar to the two-reviewer model used in standard systematic reviews. Future research could also investigate the use of LLMs to identify duplicates among the search results.

A key strength of this study is its role as proof-of-concept. It demonstrates that commercial web-based tools such as ChatGPT and Copilot can screen vast volumes of records without specialist programming expertise, suggesting a potentially feasible integration strategy into an evidence synthesis workflow. By testing varying sizes of labeled datasets, the study provides valuable insights to inform future processes. While LLMs are not flawless, they show promise as supportive tools that can assist human reviewers, significantly reducing the time required for screening. This approach could help sustain PEDro as a

freely accessible resource for users worldwide, including those in lower-middle income countries who may face limited access to high-quality physiotherapy research.

We acknowledge that this study has several limitations. First, commercial, web-based LLM tools are opaque at the architectural level, and their performance may not be entirely stable. This underscores the importance of promoting transparent reporting of model versions and timing of analyses and of using API approaches so that the findings can be comparable to other existing literature studies. Second, evaluation metrics depend on the reference standard and were used as in-context learning. The reference standard may involve a degree of human decision-making error, and the use of in-context learning may inadvertently mean that misleading information may be provided, which may subsequently influence the performance metrics. However, this standard is considered robust and follows established systematic review methods [20]. Finally, the prompts used are tailored to this proof-of-concept study, and prompt wording may influence LLM output [36], and thus, the findings may not generalize other disciplines. However, the approach using both prompt and in-context learning outlined in this study demonstrates the potential for LLMs to scale and be applied to evidence screening in other databases.

5. Conclusion

This study demonstrates the potential of commercial, web-based LLM tools to support title and abstract screening and improve efficiency of evidence syntheses. Alternative approaches to automation, such as machine learning and natural language processing, could achieve screening performance similar to or slightly higher than that of commercial tools, but they require a series of preprocessing steps.

Declaration of generative AI and AI-assisted technologies in the writing process

This study evaluated the screening ability of generative AI tools. The generative AI tools (CoPilot and ChatGPT) were employed solely for research purposes to assess their performance and capabilities in screening tasks. The author team declares that AI tools are never used as a substitute for human critical thinking, expertise, and evaluation, according to the author guidelines. The authors take full responsibility for the content of the published article.

CRediT authorship contribution statement

Wing S. Kwok: Writing − review & editing, Writing − original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Geraldine Wallbank:** Writing − review & editing, Project administration, Methodology, Investigation, Data curation, Conceptualization. **Philip Hodgson:** Writing − review & editing, Project administration, Methodology, Investigation, Data curation. **Thomas Schrader:** Data curation, Formal analysis, Investigation, Project administration, Software, Validation, Writing − review & editing. **Lexuan Shao:** Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Writing − review & editing. **Mark Elkins:** Writing − review & editing, Visualization, Methodology, Conceptualization. **Junior Fandim:** Writing − review & editing, Project administration, Investigation, Data curation. **Julia Scott:** Writing − review & editing, Project administration, Data curation. **Catherine Sherrington:** Writing − review & editing, Supervision, Methodology, Conceptualization. **Adrian C. Traeger:** Writing − review & editing, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The author team has nothing to declare.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2026.112309.

Data availability

We provided all information on Open Science Framework and Github.

References

[1] Chai KEK, Lines RLJ, Gucciardi DF, Ng L. Research screener: a machine learning tool to semi-automate abstract screening for systematic reviews. Syst Rev 2021;10(1):93. https://doi.org/10.1186/s13643-021-01635-3.

[2] Physiotherapy Evidence Database (PEDro. Physiotherapy evidence database (PEDro) statistics. Available at: https://pedro.org.au/english/about/pedro-statistics/. Accessed April 2, 2026.

[3] Physiotherapy Evidence Database (PEDro. Physiotherapy evidence database (PEDro). Available at: https://pedro.org.au/. Accessed April 3, 2026.

[4] Naseem U, Razzak I, Khan SK, Prasad M. A comprehensive survey on word representation models: from classical to state-of-the-art word representation language models. ACM Trans Asian Low Resource Lang Inf Process 2021;20(5):1−35. https://doi.org/10.1145/3434237.

[5] Alsentzer E, Murphy JR, Boag W. Publicly Available Clinical BERT Embeddings. Ithaca: Cornell University Library; 2019.

[6] Ling Y. Bio+Clinical BERT, BERT Base, and CNN Performance Comparison for Predicting Drug-Review Satisfaction. Ithaca: Cornell University Library; 2023.

[7] Agarwal S, Kweh QL, Jamali D, Wider W, Hossain SFA, Fauzi MA. How does artificial intelligence shape the productivity and quality of research in business studies? A systematic literature review and future research framework. Discover Sustain 2025;6(1):718−30. https://doi.org/10.1007/s43621-025-01480-7.

[8] Busch F, Hoffmann L, Rueger C, van Dijk EH, Kader R, Ortiz-Prado E, et al. Current applications and challenges in large language models for patient care: a systematic review. Commun Med 2025; 5(1):26. https://doi.org/10.1038/s43856-024-00717-2.

[9] Lieberum J-L, Toews M, Metzendorf M-I, Heilmeyer F, Siemens W, Haverkamp C, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use; a scoping review. J Clin Epidemiol 2025;181:111746. https://doi.org/10.1016/j.jclinepi.2025.111746.

[10] Shahab O, El Kurdi B, Shaukat A, Nadkarni G, Soroush A. Large language models: a primer and gastroenterology applications. Ther Adv Gastroenterol 2024;17:17562848241227031. https://doi.org/10.1177/17562848241227031.

[11] Zuo C, Yang X, Errickson J, Li J, Hong Y, Wang R. AI-assisted evidence screening method for systematic reviews in environmental research: integrating ChatGPT with domain knowledge. Environ Evid 2025;14(1):5. https://doi.org/10.1186/s13750-025-00358-5.

[12] Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. J Med Internet Res 2024;26:e48996. https://doi.org/10.2196/48996.

[13] Nitturi V, Flores A, Bauer DF. Using natural language processing to automate screening of abstracts for neurosurgical guideline creation. Neurosurgery 2025;97(3):736−41. https://doi.org/10.1227/neu.0000000000003450.

[14] Oami T, Okada Y, Nakada T-A. GPT-3.5 turbo and GPT-4 turbo in title and abstract screening for systematic reviews. JMIR Med Inform 2025;13:e64682. https://doi.org/10.2196/64682.

[15] Wilkins D. Automated title and abstract screening for scoping reviews using the GPT-4 Large Language Model. Ithaca: Cornell University Library; 2023.

[16] Li M, Sun J, Tan X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. Syst Rev 2024;13(1):219. https://doi.org/10.1186/s13643-024-02609-x.

[17] Matsui K, Utsumi T, Aoki Y, Maruki T, Takeshima M, Takaesu Y. Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: 3-layer strategy using GPT-3.5 and GPT-4 for systematic reviews. J Med Internet Res 2024;26:e52758. https://doi.org/10.2196/52758.

[18] Delgado-Chaves FM, Jennings MJ, Atalaia A, Wolff J, Horvath R, Mamdouh ZM, et al. Transforming literature screening: the emerging role of large language models in systematic reviews. Proc Natl Acad Sci U S A 2025;122(2):e2411962122. https://doi.org/10.1073/pnas.2411962122.

[19] Luo X, Tham YC, Giuffrè M, Ranisch R, Daher M, Lam K, et al. Reporting guideline for the use of generative artificial intelligence tools in MEdical research: the GAMER statement. BMJ Evid Based Med 2025;30(6):390−400. https://doi.org/10.1136/bmjebm-2025-113825.

[20] The Cochrane Collaboration. Cochrane Handbook Chapter 4: Searching for and selecting studies. 2026. Available at: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-04. Accessed 2 April 2026.

[21] Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. J Biomed Inform 2014;51:242−53. https://doi.org/10.1016/j.jbi.2014.06.005.

[22] Microsoft. Copilot (GPT-4) [Large Language Model]. 2025. Available at: https://copilot.microsoft.com/. Accessed 20 May 2026.

[23] Patton S. What's new in microsoft 365 copilot. Microsoft tech community. Available at: https://techcommunity.microsoft.com/blog/microsoft365copilotblog/what%e2%80%99s-new-in-microsoft-365-copilot–may-2025/4414313. Accessed November 8, 2025.

[24] OpenAI. ChatGPT [Large Language Model]. Available at: https://chat.openai.com/chat. Accessed November 8, 2025.

[25] Lewis P, Perez E, Piktus A. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Ithaca: Cornell University Library; 2021.

[26] Mukhopadhyay S, Samanta P. Feature engineering and supervised learning. Advanced data analytics using python. Berkeley, CA: Apress; 2021. https://doi.org/10.1007/978-1-4842-8005-8_3.

[27] Guido R, Ferrisi S, Lofaro D, Conforti D. An overview on the advancements of support vector machine models in healthcare applications: a review. Information (Basel) 2024;15(4):235. https://doi.org/10.3390/info15040235.

[28] Hu Y, Zhang X, Slavin V, Belsti Y, Tiruneh SA, Callander E, et al. Beyond comparing machine learning and logistic regression in clinical prediction modelling: shifting from model debate to data quality. J Med Internet Res 2025;27(3):e77721. https://doi.org/10.2196/77721.

[29] Gardazi NM, Daud A, Malik MK, Bukhari A, Alsahfi T, Alshemaimri B. BERT applications in natural language processing: a review. Artif Intell Rev 2025;58(6):166. https://doi.org/10.1007/s10462-025-11162-5.

[30] Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. Ithaca: Cornell University Library; 2020.

[31] ICF. Document classification and topic extraction resource. Available at: https://www.icf-docter.com/login. Accessed November 8, 2025.

[32] Varghese A, Cawley M, Hong T. Supervised clustering for automated document classification and prioritization: a case study using toxicological abstracts. Environ Syst Decisions 2018;38(3):398−414. https://doi.org/10.1007/s10669-017-9670-5.

[33] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval. Cambridge: Cambridge University Press; 2008.

[34] The EndNote Team. EndNote (Version X9) [Computer software]. Philadelphia, PA: Clarivate; 2024.

[35] StataCorp LLC. kappa—Interrater agreement. Available at: https://www.stata.com/manuals/rkappa.pdf. Accessed April 3, 2026.

[36] Wang L, Chen X, Deng X, Wen H, You M, Liu W, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. NPJ Digital Med 2024;7(1):41. https://doi.org/10.1038/s41746-024-01029-4.