Journal of Health Informatics

J Health Inform. 2026;18:e1841
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1841

ARTIGO COMPLETO
CBIS 2026

Accuracy of GPT-3.5-turbo in automated screening of systematic reviews on thermography in health

Acurácia do GPT-3.5-turbo na triagem automatizada de revisões sistemáticas sobre termografia em saúde

Gabriela do Manco Machado[1], Lina Alejandra Parra Cifuentes[1], Gabriel Nicolaz Nagaoka[2], Ana Clara Gomes da Silva[3], Karla Amorim Sancho[3], Maíra Araújo de Santana[3] ✉, Wellington Pinheiro dos Santos[3], Mariana Minatel Braga[1], Fausto Medeiros Mendes[1], Edgard Michel Crosato[1]

ORCID: 0000-0003-1269-5113 · 0000-0002-9157-0791 · 0000-0002-4972-4187 · 0000-0002-2823-5763 · 0000-0003-3437-2737 · 0000-0002-1796-7862 · 0000-0003-2558-6602 · 0000-0002-4469-6500 · 0000-0003-1711-4103 · 0000-0001-8559-9769

[1] Universidade de São Paulo, São Paulo (SP), Brasil.
[2] Pesquisador independente, São Paulo (SP), Brasil.
[3] Universidade Federal de Pernambuco, Recife (PE), Brasil.

ARTICLE INFORMATION

TYPE
Artigo completo · CBIS 2026

IDENTIFIER
e1841

DOI
10.59681/2175-4411.v18.iEspecial.2026.1841

RECEIVED
17 May 2026

ACCEPTED
20 July 2026

REVIEW PROCESS
Double Blind Review

HANDLING EDITOR
Juliano de Souza Gaspar

ACCESS
Open, CC BY 4.0

CORRESPONDING AUTHOR
Maíra Araújo de Santana
maira.araujosantana@ufpe.br

LICENSE

This is an open access article published under the Creative Commons CC BY 4.0 license, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited.

CC BY 4.0

ABSTRACT

Objective: To assess GPT-3.5-turbo's diagnostic accuracy for zero-shot title and abstract screening in a systematic review on thermography in health. **Method:** From 2,805 processed articles, a stratified sample of 350 was compared against two human reviewers' consensus. Performance metrics and Cohen's Kappa were calculated using a confusion matrix. **Results:** Sensitivity 0.77, specificity 0.83, positive predictive value 0.15, negative predictive value 0.99, accuracy 0.83, and balanced accuracy 0.80. Cohen's Kappa ranged 0.16-0.23, indicating slight-to-fair agreement. False negative and positive rates were 0.23 and 0.17, respectively. **Conclusion:** Even in zero-shot configuration and implemented by non-specialists, GPT-3.5-turbo shows potential for automating systematic review screening. Its high recall and negative predictive value enable safe initial triage, while low precision underscores the continued necessity of human oversight for final inclusion decisions.

Keywords: GPT-3.5-turbo; data mining; machine learning.

RESUMO

Objetivo: Avaliar a precisão diagnóstica do GPT-3.5-turbo na triagem zero-shot de títulos e resumos para revisão sistemática sobre termografia na saúde. **Método:** De 2.805 artigos, uma amostra estratificada de 350 foi comparada ao consenso de dois revisores humanos. Métricas de desempenho e Kappa de Cohen foram calculados via matriz de confusão. **Resultados:** Sensibilidade 0,77, especificidade 0,83, valor preditivo positivo 0,15, valor preditivo negativo 0,99, acurácia 0,83 e balanceada 0,80. Kappa variou de 0,16 a 0,23 (concordância fraca a regular). Taxas de erro: falso negativo 0,23 e falso positivo 0,17. **Conclusão:** Mesmo em configuração zero-shot e por equipe não especializada, o modelo apresenta potencial para automatizar triagem em revisões sistemáticas. Alto recall e valor preditivo negativo permitem triagem inicial segura, enquanto baixa precisão mantém a necessidade de supervisão humana para decisões finais de inclusão.

Descritores: GPT-3.5-turbo; mineração de dados; aprendizado de máquina.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Paper presented at the 21st Brazilian Congress of Health Informatics (CBIS 2026), held from September 23 to 25, 2026, in Brasília (DF), Brazil, and published in the CBIS 2026 section of the Journal of Health Informatics.

HOW TO CITE THIS ARTICLE: Machado GM, Cifuentes LAP, Nagaoka GN, da Silva ACG, Sancho KA, de Santana MA, et al. Accuracy of GPT-3.5-turbo in automated screening of systematic reviews on thermography in health. J Health Inform 2026;18:e1841. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1841

1 INTRODUCTION

Systematic reviews (SRs) synthesize primary studies that address a defined research question. In Health, SRs summarize findings and provide a clinical recommendation based on evidence certainty, thereby supporting informed decision-making. However, the number of available studies has been increasing over time, making some stages of systematic reviews particularly challenging, especially the initial stage of title and abstract screening [1,2], when researchers must decide which papers are truly relevant to the research question and deserve full-text assessment.

In addition to the considerable workload, which increases the time and resources required to complete the task, the cognitive burden [3,4] also could raise the likelihood of bias. The screening phase, usually based on titles and abstracts, is particularly prone to anchoring bias: for example, reviewers may give disproportionate weight to the first term or phrase that seems relevant, making them reluctant to revise an early impression [5,6]. Reviewers also tend to rely on confirmation and availability heuristics, giving preference to records that match their expectations or come most readily to mind. At the same time, framing effects in abstracts, as well as premature closure and overconfidence, can lead them to stop considering alternative options too early [5,7].

Efforts to automate or simplify these labor-intensive stages of systematic reviews are not new. Over the years, researchers have explored various approaches to reduce the workload and cognitive burden associated with screening and data extraction [1,2]. Early attempts focused on semi-automated tools and rule-based algorithms to assist human reviewers, while more recent developments have leveraged advances in natural language processing and machine learning to prioritize or classify studies more efficiently.

Nevertheless, Large Language Models (LLMs) and other computational algorithms may also introduce biases, such as overweighting high-frequency tokens or drifting during sampling [8,9]. Even so, they offer a level of automation that can democratize access to screening tools. Early studies in radiology reported that ChatGPT achieved at least 95% sensitivity and 99% negative predictive value, although its precision remained lower than that of human reviewers [10]. A methodological review of 56 evaluations confirmed consistently high recall but noted an increased number of false positives and considerable variability across topics [11]. Large-scale experiments on Cochrane data (approximately 120,000 records) employed optimized prompting and a human–LLM hybrid workflow to preserve full sensitivity [12]. Similarly, embedding GPT-4 into a review on vitamin D and falls reduced manual screening time by 95% without missing any eligible studies [13].

These results look promising, but researchers still need to learn how review teams without artificial intelligence (AI) expertise perform when deploying LLMs. Closing this knowledge gap could democratize adoption and broaden the evidence base. We therefore evaluated GPT-3.5-turbo in zero-shot mode during the screening phase of an ongoing SR led by health researchers with no previous LLM experience. Every AI screening decision was reviewed and the resulting human consensus served as the gold standard, offering a realistic measure of what LLMs can and cannot deliver at the most labor-intensive stage of evidence synthesis.

2 METHODS

In this manuscript, we focus on the screening phase of a systematic review on thermography in health, evaluating how an LLM classifies titles and abstracts against the consensus of human researchers as the reference standard.

We conducted a systematic review on the accuracy of AI-enhanced thermography in health (PROSPERO: CRD42024551259). All data related to this manuscript and the SR will be made available after the final manuscript is published through an institutional repository. We searched PubMed, Cochrane Library, Digital Library, Scopus, Web of Science, arXiv, IEEE Xplore, ACM, and MEDLINE (Additional File 1) and retrieved 2,805 articles after removing duplicates with Rayyan software (Rayyan Systems Inc., https://rayyan.ai).

Sampling

This diagnostic accuracy study employed a cross-sectional design. A pilot screening of the first 761 records, conducted by two human reviewers, indicated an estimated prevalence of 21% of eligible studies. We randomly sampled 350 articles to achieve a ±7% margin of error at a 95% confidence level. Using the pandas library in Python with random_state=42, we ensured reproducibility. We then screened the selected records (n=350) by titles and abstracts according to predefined eligibility criteria, with two human reviewers and an LLM algorithm assessing them independently.

2.1 Eligibility criteria and human researchers screening

We instructed both the human reviewers and the LLM with predefined eligibility criteria. **Inclusion criteria:** studies evaluating the diagnostic accuracy of AI-assisted thermography in health applications. Eligible designs included observational (cohort and case-control), cross-sectional, and randomized clinical trials, as long as they reported at least one metric: sensitivity, specificity, positive predictive value (PPV), or negative predictive value (NPV). **Exclusion criteria:** studies with incomplete or unpublished data, inadequate methodology (e.g., missing

Table 1. Performance metrics (outcomes)

Outcomes	Metric	Formula
Primary	Accuracy	(TP + TN) / (TP + TN + FP + FN)
Secondaries	Sensitivity	TP / (TP + FN)
	Specificity	TN / (TN + FP)
	Positive Predictive Value	TP / (TP + FP)
	Negative Predictive Value	TN / (TN + FN)
	Balanced Accuracy	(Sensitivity + Specificity) / 2
	False Negative Rate	FN / (FN + TP)
	False Positive Rate	FP / (FP + TN)

TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives

Source: Prepared by the authors (2026).

details on thermal image resolution), or those using positron emission tomography (PET) or outdated thermographic techniques (e.g., printed-paper thermography or low-resolution thermal imaging).

Two independent human reviewers screened the titles and abstracts of the sample (n=350), without access to each other's decisions or to the LLM's output (blinded). We classified each record as 1 (included) or 0 (excluded) according to the predefined eligibility criteria. We carried out the screening in an Excel spreadsheet after a calibration process. When disagreements arose, the reviewers resolved them through discussion and reached consensus, which we used as the reference standard to evaluate the LLM's performance.

2.2 Large Language Model Algorithm

An LLM was implemented using the OpenAI API (GPT-3.5-turbo) to assist in the study screening process. The model received the following prompt:

prompt = f"""

You are a researcher participating in a SR. Your job is to screen studies and decide upon the inclusion and exclusion criteria below, if they should be included in the review.

Make your decision only based on the given criteria, title and abstract. Answer it with 'True' if the study should be included, and 'False' otherwise.

"""

Although our analysis focuses on a sample of 350 records, the LLM screened all 2,805 records retrieved from the literature search, based solely on titles and abstracts without access to the human reviewers' decisions. We used the model in a zero-shot setting, relying only on the prompt and applying no task-specific fine-tuning or prior training. The LLM returned "True" for included and "False" for excluded studies, and these outputs were converted to 1 and 0, respectively, to align with the human reviewers' classification.

2.3 Outcome Measures

We chose accuracy as the primary outcome because it provides an overall measure of the model's performance and guided the sample size estimation. We defined accuracy as the proportion of correctly classified cases, calculated by dividing the sum of true positives and true negatives by the total number of records. However, due to the class imbalance typically found in screening scenarios, additional metrics such as sensitivity and negative predictive value (NPV) were also prioritized in the interpretation of the results.

2.4 Statistical Analysis

We preprocessed all classification data with the pandas library to ensure consistency and standardize formatting. A confusion matrix was generated by comparing the binary classification decisions made by the GPT-3.5-turbo model (1 = include, 0 = exclude) with the final human consensus labels.

We obtained TP, TN, FP, and FN using the confusion_matrix function from the *sklearn.metrics* module. We used these values to compute all performance metrics described above. We calculated Cohen's kappa coefficient using the *cohen_kappa_score* function from sklearn.metrics, to evaluate the agreement between the LLM's classifications and the human reviewers' consensus.

We also qualitatively assessed the disagreements between the LLM and the human consensus to understand what may have caused the mismatches. This includes checking if there were issues like unclear abstracts, borderline cases, or anything in the way the prompt was interpreted. This allowed us to identify possible patterns and better understand the model's limitations.

3 RESULTS

Table 2 shows the confusion matrix comparing the LLM's decisions against the dual-reviewer consensus gold standard. The model

Table 2. Confusion matrix comparing GPT-3.5-turbo and dual human reviewer consensus. Shows true positives (n=10), false positives (n=57), false negatives (n=3), and true negatives (n=280) from the title-and-abstract screening phase.

LLM Prediction / Human Consensus	Included (Positive)	Excluded (Negative)	Total
Included (Positive)	10 (TP)	57 (FP)	67
Excluded (Negative)	3 (FN)	280 (TN)	283
Total	13	337	350

Source: Prepared by the authors (2026).

Table 3. Performance metrics of GPT-3.5-turbo in the screening phase

Outcome	Metric	Value
Primary	Accuracy	0.83
Secondaries	Sensitivity	0.77
	Specificity	0.83
	Positive Predictive Value	0.15
	Negative Predictive Value	0.99
	Balanced Accuracy	0.80
	False Negative Rate	0.23
	False Positive Rate	0.17

Source: Prepared by the authors (2026).

correctly excluded 280 records (true negatives) and produced three false negatives. While the absolute number of missed inclusions was low (n=3), this corresponds to a false negative rate of 23% (sensitivity of 0.77), indicating that the model missed nearly a quarter of the eligible studies in the sample (Table 2).

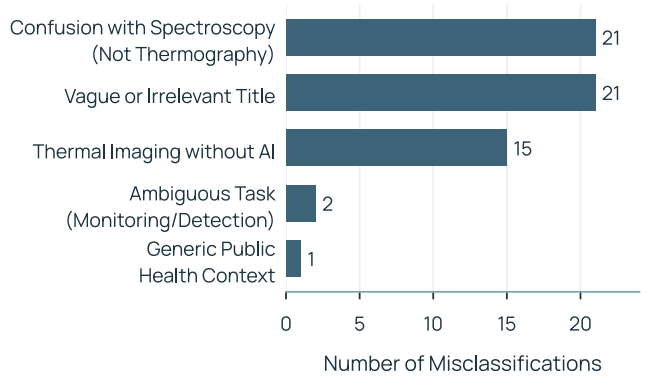
We calculated Cohen's kappa coefficient to assess the agreement between the LLM and each human reviewer individually. The results indicated slight agreement between the model and Reviewer 1 (kappa of 0.16) and fair agreement with Reviewer 2 (kappa of 0.23).

We reviewed final disagreements between the LLM and the human consensus jointly by Reviewers 1 and 2 to qualitatively explore potential sources of error. A total of 60 records (57 false positives and 3 false negatives) were classified differently by the LLM compared to the human consensus (Additional file 2). Human reviewers reassessed all disagreements to qualitatively explore common sources of error. We qualitatively reviewed all misclassifications and grouped them into thematic categories. As shown in Figure 1, the most frequent causes of error were vague or ambiguous titles (n=21), confusion with non-thermal imaging techniques such as spectroscopy or infrared analysis (n=21), and studies involving thermal imaging but without the use of artificial intelligence (n=15). Less frequent categories included general public health contexts and ambiguous tasks such as "monitoring" or "detection" without clear diagnostic intent.

4 DISCUSSION

The use of LLMs is emerging as a promising approach for automating parts of evidence synthesis under human supervision, particularly in early screening stages where recall is prioritized. Compared with human consensus, the model retrieved 77% of eligible papers, resulting in a false negative rate of 23%. Although the absolute number of missed studies was small (n=3), this sensitivity level indicates that the model cannot replace human screening. The model achieved a negative predictive value (NPV) of 0.99; however, it

Figure 1. Thematic distribution of screening errors by GPT-3.5-turbo. Percentage of disagreement instances categorized by error type (e.g., vague abstracts, out-of-scope studies, absence of AI mention)



Source: Prepared by the authors (2026).

is important to note that NPV is highly dependent on the prevalence of eligible studies in the sample. In scenarios with lower prevalence, NPV would naturally increase, but the false negative rate would remain the critical safety metric. The balanced accuracy reached 0.80, while precision remained modest at 0.15, reflecting the inclusion of irrelevant studies, which could be removed later. This trade-off between

moderate recall and lower precision aligns with previous findings in the literature, suggesting that LLMs can substantially reduce human workload [10-13]. Nevertheless, the priority remains to ensure that no relevant study is missed.

Earlier studies have mapped the trade-off between recall and workload. In radiology, ChatGPT reached 95% sensitivity and 99% NPV but showed lower precision than human reviewers [10]. Our strategy also showed low precision (0.15), meaning that the model favored inclusion to minimize the risk of excluding potentially relevant studies (false negatives). A recent methodological review reported that precision tended to vary across topics, despite consistently high recall [11]. Large-scale Cochrane experiments demonstrated that GPT-4o and Gemini 1.5 Pro could achieve full sensitivity when prompts were optimized, though at the cost of additional false positives [12]. Similarly, a prospective review on vitamin D and falls reduced manual screening effort by 95% without missing any eligible records [13]. In relation to accuracy, although the overall accuracy appeared acceptable (0.83), this metric can be misleading in this context, since literature searches typically retrieve a large number of obviously irrelevant titles, which the model can easily exclude. As a result, accuracy may appear artificially high, since correctly rejecting trivial negatives contributes disproportionately to the overall score. This inflation can give a misleading impression of model performance, emphasizing the need to prioritize more informative metrics such as sensitivity (catching as many relevant studies as possible) and negative predictive value (how safe it is to discard a study).

Like other recent studies [10,11,13], our approach relied on a pre-configured model and was not iteratively refined. This design choice reflects how LLMs are currently being applied in evidence synthesis by health sciences researchers using ready-to-use models to assess feasibility and operational safety in real-world review settings, rather than to optimize performance. What distinguishes our study is the deliberate evaluation of a widely accessible, lower-cost LLM to determine whether acceptable safety could be achieved under realistic constraints faced by many research teams. Despite its lower capacity compared with newer models, GPT-3.5-turbo achieved performance that is promising for a first-pass triage strategy. Although newer engines demonstrate higher balanced accuracy and precision [10,11], our results indicate that a useful baseline of screening performance can be achieved with more widely accessible, lower-cost models, provided that human oversight is maintained.

Another factor that may have influenced performance, and may constrain external validity, is the highly jargon-dense nature of the field [11,14]. While this is an inherent characteristic, researchers applying LLMs in similarly technical domains should carefully define their eligibility criteria, as specialized terminology and abbreviations can add complexity [11]. Allowing the model to analyze full-text articles rather than only titles and abstracts could increase both precision and recall without additional runtime for the models [13], representing a straightforward next step for improvement. We limited our analysis to title-and-abstract screening to align with standard human workflows and ensure a fair comparison.

While model performance can be influenced by technical factors such as prompt design or field-specific terminology, human reviewers are also prone to limitations that affect decision consistency during screening. Interestingly, before consensus was reached, the agreement between the model and each reviewer was only slight for Reviewer 1 ($\kappa = 0.16$) and fair for Reviewer 2 ($\kappa = 0.23$). This low Cohen's kappa reflects not only the differences in how the LLM and human reviewers interpret the eligibility criteria, but also the inherent subjectivity in screening highly technical abstracts. While humans may infer context from ambiguous phrasing, the LLM strictly adheres to the literal text of the prompt. Most disagreements seem to arise from differences in how the reviewers interpreted and applied the eligibility criteria. Reviewer 1 prioritized the inclusion of studies that explicitly combined thermography with artificial intelligence techniques, whereas Reviewer 2 gave greater weight to any diagnostic use of thermography, even when the AI component was not clearly stated. These differences in emphasis led to minor discrepancies that were quickly resolved by consensus. However, such variations illustrate the influence of cognitive biases that may affect the screening phase and, consequently, introduce bias into systematic reviews, such as decision fatigue after prolonged screening sessions or anchoring and confirmation bias when reviewers adhere too rigidly to early impressions or predefined rules [5-7]. LLMs, however, are subject to different sources of bias. While they do not experience fatigue or anchoring, their decisions are rigidly bound to prompt design and textual formulation. A slight change in the prompt can shift recall and precision by double-digit margins, and an abstract that phrases a concept in an unusual way may be mistakenly excluded as a false negative. Because the model does not learn or adapt mid-run, it cannot correct these misclassifications on its own [8,9]. Auditing the 60 disagreements (vague abstracts, non-thermal imaging, AI-free thermal studies) offers clear targets for prompt refinement.

Combining computational and human approaches currently appears to be the most effective strategy. Manual title-and-abstract screening is time-consuming, typically requiring 45 to 60 seconds per record according to previous studies [1,2]. This workload scales rapidly with the number of retrieved citations, making automation particularly appealing for large-

scale evidence syntheses. In contrast, LLM-based screening can complete the same task in a matter of seconds [10,13], although the exact duration depends on how the process is structured: whether abstracts are evaluated individually, in batches, or all at once. In our study, reviewers still had to verify 57 false positives (17%), but filtering out 280 clearly irrelevant records saved several hours of manual effort. In this context, maintaining high recall is more valuable than achieving perfect precision [12,15-17].

5 CONCLUSIONS

GPT-3.5-turbo can provide value as a first-pass triage tool in title-and-abstract screening when high recall is prioritized and human oversight is maintained. A notable limitation of this study, however, is its restriction to the GPT-3.5-turbo model, which is now considered a legacy version in the rapidly evolving landscape of Large Language Models. Future research should leverage newer, more efficient, and cost-effective models (e.g., GPT-4o-mini), which have the potential to yield even better performance at lower computational costs. Furthermore, while overall accuracy was utilized in this study to guide sample size estimation, future evaluations of LLMs in systematic review screening should prioritize more informative metrics for highly imbalanced datasets, such as sensitivity, false negative rates, and negative predictive value, rather than relying on accuracy as the primary outcome. Finally, future work should target precision gains through prompt ensembles and retrieval-augmented strategies, while assessing whether these improvements justify their computational and energy costs.

Author contributions. The authors contributed equally to all stages of the article: conception and design of the study; data collection, analysis, and interpretation; drafting and critical revision of the manuscript; and approval of the final version.

Acknowledgments. The authors acknowledge the Mamo.IA research group for the continuous exchange of knowledge and interdisciplinary collaboration that supported the development of this work.

Funding. This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq) through a DTI-A scholarship (Grant No. 383046/2024-6; Proposal No. 9646049472643444), as part of the project "Mamo.IA: Development of an Artificial Intelligence Platform for Early Diagnosis of Breast Cancer with Emphasis on Mammography, Immunohistochemistry, and Thermography" (Process No. 445896/2023-0), funded under Call No. 21/2023, Track B (Primary and Original Transdisciplinary Studies).

Data and materials availability. The authors did not provide a data and materials availability statement.

Conflicts of interest. The authors do not have any conflicts of interest to declare.

Use of artificial intelligence. In the methodological stage of title and abstract screening, the GPT-3.5-turbo model (OpenAI) was used, as detailed in the Methods section. Additionally, the Qwen3.6-Plus language model was employed exclusively for orthographic and grammatical editing of the manuscript. No artificial intelligence was involved in the conception of the research, methodological design, data collection or analysis, interpretation of results, or generation of intellectual content. All AI-assisted processes and outputs were rigorously validated and audited by the human authors, who assume full responsibility for the accuracy, originality, and scientific integrity of this work.

REFERENCES

1. 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. doi: https://doi.org/10.1016/j.jbi.2014.06.005
2. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics. 2010;11:55. doi: https://doi.org/10.1186/1471-2105-11-55
3. Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol. 2020;121:20-8. doi: https://doi.org/10.1016/j.jclinepi.2020.01.005
4. Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med Res Methodol. 2019;19(1):132. doi: https://doi.org/10.1186/s12874-019-0782-0
5. Polanin JR, Pigott TD, Espelage DL, Grotpeter JK. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. Research Synthesis Methods. 2019;10(3):330-42. doi: https://doi.org/10.1002/jrsm.1354
6. Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PLoS One. 2020;15(1):e0227742. doi: https://doi.org/10.1371/journal.pone.0227742
7. Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak. 2016;16(1):138. doi: https://doi.org/10.1186/s12911-016-0377-1
8. Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. J Am Med Inform Assoc. 2024;31(5):1172-83. doi: https://doi.org/10.1093/jamia/ocae060

9. Mangat G, Sharma S, Bergemann R. HTA101 Artificial Intelligence Bias in Systematic Literature Reviews (SLRS) for Health Technology Assessment (HTA). Value in Health. 2023;26(6):S276-S277. doi: https://doi.org/10.1016/j.jval.2023.03.2667

10. Issaiy M, Ghanaati H, Kolahi S, Shakiba M, Jalali AH, Zarei D, et al. Methodological insights into ChatGPT's screening performance in systematic reviews. BMC Med Res Methodol. 2024;24(1):78. doi: https://doi.org/10.1186/s12874-024-02203-8

11. Galli C, Gavrilova AV, Calciolari E. Large Language Models in Systematic Review Screening: Opportunities, Challenges, and Methodological Considerations. Information. 2025;16(5):378. doi: https://doi.org/10.3390/info16050378

12. Sanghera R, Thirunavukarasu AJ, El Khoury M, O'Logbon J, Chen Y, Watt A, et al. High-performance automated abstract screening with large language model ensembles. J Am Med Inform Assoc. 2025;32(5):893-904. doi: https://doi.org/10.1093/jamia/ocaf050

13. Trad F, Yammine R, Charafeddine J, Chakhtoura M, Rahme M, Fuleihan GE-H, et al. Streamlining Systematic Reviews: A Novel Application of Large Language Models. arXiv [Preprint]. 2024 [citado em 16 abr 2026]. Disponível em: https://arxiv.org/abs/2412.15247.

14. Mai HT, Chu CX, Paulheim H. Do LLMs Really Adapt to Domains? An Ontology Learning Perspective. In: Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2025. p. 126-43. doi: https://doi.org/10.1007/978-3-031-77844-5_7

15. Adam GP, DeYoung J, Paul A, Saldanha IJ, Balk EM, Trikalinos TA, et al. Literature search sandbox: a large language model that generates search queries for systematic reviews. JAMIA Open. 2024;7(3):ooae098. doi: https://doi.org/10.1093/jamiaopen/ooae098

16. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):163. doi: https://doi.org/10.1186/s13643-019-1074-9

17. Stansfield C, O'Mara-Eves A, Thomas J. Text mining for search term development in systematic reviewing: A discussion of some methods and challenges. Res Synth Methods. 2017;8(3):355-65. doi: https://doi.org/10.1002/jrsm.1250

EDITORIAL NOTES

The Journal of Health Informatics is a scientific journal maintained by the Brazilian Society of Health Informatics (Sociedade Brasileira de Informática em Saúde), a non-profit scientific association dedicated to the development of Health Informatics and Digital Health in Brazil. The editorial process is conducted with scientific independence, free from institutional, commercial, or sponsor interference, and follows international best practices in publication ethics.

This article was double-blind peer reviewed by specialists in the field. The editorial team used artificial intelligence tools, under human supervision, in the initial screening, in checking the text, references and numbers and, when needed, in redrawing figures, charts and tables, which the authors approved in the proof. The tools and safeguards are described in the journal's Policy on the use of artificial intelligence. Editorial decisions were made by people.

SBIS Sociedade Brasileira de Informática em Saúde