Educitec
Revista de Estudos e Pesquisas sobre Ensino Tecnológico
Revista do Programa de Pós-Graduação em Ensino Tecnológico, do Instituto Federal do Amazonas
ISSN: 2446-774X

Humanos vs. inteligências artificiais em revisões sistemáticas: avaliação nas etapas de seleção por títulos e resumos

Fernando Emilio Puntel[1]
Muriel Belo Pereira[2]
Bruna Adriane Fary[3]
Gerson Geraldo Homrich Cavalheiro[4]

Resumo
A Inteligência Artificial (IA) tem se consolidado como ferramenta relevante na pesquisa científica, especialmente no apoio a tarefas analíticas e repetitivas que demandam elevado esforço humano. No contexto da pesquisa educacional, as Revisões Sistemáticas de Literatura (RSL) representam processos complexos e demorados, sobretudo nas etapas iniciais de triagem. Este estudo analisa a eficácia de duas IAs generativas em suas versões gratuitas (ChatGPT e Gemini) no apoio às fases iniciais de uma RSL sobre a interseção entre Pensamento Computacional, Química e STEM/STEAM. Dois revisores humanos analisaram, de forma independente, os títulos e resumos de 1.051 artigos. Paralelamente, as duas IAs realizaram a mesma triagem, sob supervisão de um terceiro revisor. Na análise dos títulos, as IAs selecionaram diversos artigos com termos-chave relevantes, mas sem relação direta com o foco da pesquisa, gerando quantidade significativa de falsos positivos. Por outro lado, na análise dos resumos, as IAs demonstraram maior precisão, identificando inclusive artigos elegíveis não selecionados pelos revisores humanos. Dois desses estudos relevantes foram recuperados apenas pelas ferramentas de IA, evidenciando seu potencial como apoio complementar. Os resultados indicam que, embora apresentem limitações na triagem inicial, as IAs podem contribuir significativamente para acelerar as etapas preliminares de RSL, especialmente quando utilizadas como suporte complementar ao julgamento humano.

Palavras-chave: revisão sistemática da literatura, inteligência artificial na educação, tecnologias educacionais, ensino de Ciências.

Humans vs. artificial intelligences in systematic reviews: evaluation in the title and abstract screening stages

Abstract
Artificial Intelligence (AI) has become an increasingly relevant tool in scientific research, particularly in supporting analytical and repetitive tasks that require substantial human effort. In the context of educational research, Systematic Literature Reviews (SLRs) represent complex and time-consuming

[1] Mestre em Ciências da Computação, Santa Maria, RS, Brasil. ORCID: https://orcid.org/0000-0003-0333-288X E-mail: fepuntel@inf.ufpel.edu.br
[2] Mestre em Química, Universidade Federal de Pelotas. Pelotas, RS, Brasil. ORCID: https://orcid.org/0000-0003-3463-8195 E-mail: muriel.belo@hotmail.com
[3] Doutorado em Ensino de Ciências e Educação Matemática, Universidade Estadual de Londrina, Londrina, PR, Brasil. ORCID: https://orcid.org/0000-0002-2382-6572 E-mail: fary.bruna@gmail.com
[4] Doutor em Informatique Systèmes et Communications, Institut National Polytechique de Grenoble, INPG, França. ORCID: https://orcid.org/0000-0002-4314-3429 E-mail: gerson.cavalheiro@inf.ufpel.edu.br

processes, especially during the initial screening stages. This study analyzes the effectiveness of two generative AI tools in their free versions (ChatGPT and Gemini) in supporting the early stages of an SLR focused on the intersection of Computational Thinking, Chemistry, and STEM/STEAM. Two human reviewers independently analyzed the titles and abstracts of 1,051 articles. In parallel, the two AI tools performed the same screening under the supervision of a third reviewer. In the title screening phase, the AIs selected several articles containing relevant keywords but lacking direct alignment with the research focus, resulting in a significant number of false positives. In contrast, during abstract screening, the AIs demonstrated greater precision, identifying eligible articles that had not been selected by the human reviewers. Two of these relevant studies were retrieved exclusively by the AI tools, highlighting their potential as complementary support. The findings indicate that although AIs present limitations in the initial screening stage, they can significantly contribute to accelerating the preliminary phases of SLRs, particularly when used as complementary support to human judgment.

Keywords: systematic literature review, artificial intelligence in education, educational technologies, science education

Humanos vs. inteligencias artificiales en revisiones sistemáticas: evalución en las etapas de selección por títulos y resúmenes

Resumen
La Inteligencia Artificial (IA) se ha consolidado como una herramienta relevante en la investigación científica, especialmente en el apoyo a tareas analíticas y repetitivas que demandan un alto esfuerzo humano. En el contexto de la investigación educativa, las Revisiones Sistemáticas de Literatura (RSL) representan procesos complejos y prolongados, especialmente en las etapas iniciales de selección. Este estudio analiza la eficacia de dos herramientas de IA generativa en sus versiones gratuitas (ChatGPT y Gemini) en el apoyo a las fases iniciales de una RSL centrada en la intersección entre Pensamiento Computacional, Química y STEM/STEAM. Dos revisores humanos analizaron de forma independiente los títulos y resúmenes de 1.051 artículos. Paralelamente, las dos IAs realizaron el mismo proceso de selección bajo la supervisión de un tercer revisor. En el análisis de títulos, las IAs seleccionaron varios artículos con palabras clave relevantes, pero sin relación directa con el enfoque central de la investigación, generando una cantidad significativa de falsos positivos. Por otro lado, en el análisis de resúmenes, las IAs demostraron mayor precisión, identificando incluso artículos elegibles que no habían sido seleccionados por los revisores humanos. Dos de estos estudios relevantes fueron recuperados exclusivamente por las herramientas de IA, lo que evidencia su potencial como apoyo complementario. Los resultados indican que, aunque presentan limitaciones en la etapa inicial de selección, las IAs pueden contribuir significativamente a acelerar las fases preliminares de las RSL, especialmente cuando se utilizan como complemento al juicio humano.

Palabras clave: revisión sistemática de la literatura, inteligencia artificial en la educación, tecnologías educativas, enseñanza de las Ciencias.

Introdução

A utilização de Inteligência Artificial (IA) tem se integrado de forma crescente ao nosso cotidiano, especialmente após a popularização de ferramentas oferecendo recursos de sumarização, sintetização e interpretação de linguagem natural por *Large Language Models*, ou Modelos de Linguagem de Grande Escala (LLMs) e de produção de conteúdo por IAs generativas.

Dentre as ferramentas de IA generativas mais populares atualmente encontram-se o ChatGPT e o Google Gemini. Embora ambas possuam versões gratuitas de acesso, essas modalidades oferecem recursos limitados quando

comparadas às versões pagas, que disponibilizam modelos mais avançados e funcionalidades ampliadas. Na presente pesquisa, o uso dessas ferramentas considerou exclusivamente suas versões gratuitas, com todas as restrições quando comparadas as versões não gratuitas.

A integração da IA no contexto educacional, especialmente por meio de ferramentas de IA generativas que prometem contribuições significativas para o processo de ensino e aprendizagem, além de um olhar atento para sua utilização (Santos, Santo; 2025). Tais avanços incluem a potencialização da personalização do ensino, o apoio à formação continuada de professores e a reconfiguração das práticas docentes, além de suscitar discussões cruciais sobre privacidade de dados e equidade na implementação (Menta e Brito, 2024). Evangelista *et al.* (2023) destacaram que instituições de ensino estão enfrentando desafios com o advento das LLMs e das IAs generativas como ferramentas acessíveis aos alunos, tornando-se necessário avaliar tanto os benefícios quanto às limitações de seu uso.

O meio acadêmico tem se beneficiado em diversas áreas, desde o uso de IA para análise de dados e correções de trabalhos, o auxílio na produção de textos acadêmicos (Kacena; Plotkin; Fehrenbacher, 2024) e até benefícios na escrita acadêmica, incluindo considerações sobre plágios, preconceitos e integridade acadêmica (William, 2024). Em particular, Van Dijk (2023) destaca o papel da IA em revisões sistemáticas, onde ferramentas como o "*ASReview*" aceleram a triagem de artigos, sugerindo aqueles com maior relevância com base em aprendizado automatizado. Trad *et al.* (2025) propuseram um sistema baseado em LLM para automatizar tanto a triagem de títulos/resumos quanto a triagem de texto completo, reduzindo drasticamente o tempo de revisão sem comprometer sensibilidade e especificidade. Esses processos resultam em economia de tempo e mantém a qualidade metodológica, desde que sejam seguidos critérios rigorosos de controle e verificação entre revisores.

A grande quantidade de artigos publicados torna a execução de uma Revisão Sistemática da Literatura (RSL) um desafio que demanda organização, concentração e tempo (Borah et al., 2017). Métodos empregando LLMs e IA generativas têm se mostrado promissores para reduzir o trabalho humano durante essas revisões. Blaizot et al. (2022) conduziram uma revisão sistemática sobre ferramentas automatizadas baseadas em IA para apoio a revisões nas ciências da saúde e concluíram que,

embora esses métodos estejam em consolidação, ainda apresentam limitações quanto à confiabilidade, podendo ocorrer perda de artigos e falhas na extração de dados, o que exige intervenção e julgamento humano.

Para garantir rigor metodológico, recomenda-se a adoção de diretrizes internacionalmente reconhecidas, como o checklist Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021), além de orientações práticas descritas por Aromataris e Pearson (2014) para a condução de RSLs. Tais diretrizes reforçam a necessidade de transparência, reprodutibilidade e controle criterioso nas etapas do processo.

Considerando que uma RSL pode levar, em média, 67,3 semanas para ser concluída (Borah et al., 2017), foram desenvolvidas ferramentas baseadas em IA para apoiar etapas como triagem e extração de dados. Plataformas como DistillerSR, Rayyan, Elicit e RobotReviewer buscam facilitar esse processo; contudo, grande parte dessas ferramentas oferece funcionalidades completas apenas em versões pagas. Em contraste, ferramentas de uso geral como ChatGPT e Gemini disponibilizam versões gratuitas, embora ainda apresentem limitações quanto à precisão e personalização. Diante da complexidade das RSLs e dos debates éticos relacionados ao uso de IA na produção científica (Khalifa, Albadawy; 2024), torna-se relevante investigar o potencial dessas ferramentas como apoio supervisionado nas etapas iniciais de revisão, preservando a responsabilidade metodológica e científica do pesquisador.

Este estudo analisa a eficácia de duas ferramentas de IA generativa em suas versões gratuitas (ChatGPT (GPT-3.5) e Gemini (1.5)) no apoio às etapas iniciais de uma Revisão Sistemática da Literatura, especificamente na leitura de títulos e resumos. A aplicação foi realizada em uma temática interdisciplinar envolvendo Pensamento Computacional, Química e STEM/STEAM, utilizada como contexto para comparação, mas não como objeto central da análise. As ferramentas foram selecionadas por utilizarem arquiteturas distintas: o ChatGPT é baseado na tecnologia Generative Pre-trained Transformer (GPT), enquanto o Gemini adota um modelo multimodal de linguagem neural conversacional (Rane; Choudhary; Rane, 2024).

Como estudo de caso, é apresentado um exemplo em que a RSL é conduzida de forma híbrida, comparando e contrastando os resultados obtidos nas etapas iniciais de triagem de artigos. Essas etapas incluem a aplicação de critérios de exclusão e

inclusão, bem como a análise de títulos e resumos, realizadas por revisores humanos e pelas duas ferramentas de IA. Este estudo busca analisar o potencial das IAs em apoiar as etapas mais demoradas da construção de uma RSL, especialmente na triagem inicial de grandes volumes de artigos recuperados nas bases de dados.

Métodos

O tema abordado neste estudo é a aplicação do Pensamento Computacional (PC) na Química por meio da metodologia STEAM, área que integra a linha de pesquisa dos autores. Essa vinculação acadêmica reforça a relevância e a consistência da análise realizada, além de contribuir para a discussão de implicações práticas e perspectivas para pesquisas futuras. As bases selecionadas para busca de artigos foram: *IEEExplore*, ACM, ERIC, *PubMed*, *Wiley Online Library* e EBSCO. Entre as bases selecionadas, estão aquelas que se destacam por sua representatividade na área de Computação, bem como as que incluem artigos com temas relevantes à RSL e relacionados à Química. Para a realização das buscas, foram testadas seis diferentes *strings* de pesquisa. A *string* que apresentou o melhor resultado, retornando os artigos sentinelas, estudos previamente identificados como altamente relevantes para a temática e utilizados como referência para validar a eficácia da busca, foi: (“*Computational thinking*”) AND (”STEAM” OR “STEM”) AND (“*Chemistry*” OR “*Chemical education*”).

Todas as buscas foram realizadas no dia 28 de março de 2024, onde as informações foram salvas e somente após isso as análise começaram a ser realizadas, garantindo assim um recorte das bases de dados, e não foram atualizadas/refeitas, pois o objetivo principal deste estudo é analisar a confiabilidade da IA em comparação ao processo humano, e não obter os artigos mais recentes sobre o tema. O processo de busca nas bases de dados foi realizado por um dos autores humanos. O resultado desta busca foi exportado e organizado manualmente em planilhas do Google (Google Sheets), as quais foram utilizadas para distribuir os registros entre os dois revisores humanos e o revisor responsável pela triagem assistida por ferramentas de IA. Esse procedimento possibilitou o controle e o acompanhamento das decisões de inclusão e exclusão de forma estruturada e acessível a todos os envolvidos. As ferramentas de IA utilizadas foram o ChatGPT versão 3.5 e o Gemini, versão 1.5.

Os artigos resultantes das buscas passaram por duas etapas de triagem:

- Leitura dos títulos: os títulos foram lidos e classificados por dois revisores independentes (Autor 1 e Autor 2) e pelas IAs.
- Leitura dos resumos: os resumos completos dos artigos com títulos incluídos ou em dúvida na etapa anterior foram lidos por dois revisores independentes (Autor 1 e Autor 2) e pelas IAs.

Para garantir o rigor metodológico e minimizar vieses individuais, qualquer divergência entre os dois revisores foi resolvida por um terceiro revisor (iniciais), especialista na área de estudo. O processo de resolução de divergências seguiu um protocolo de consenso, onde os revisores discutiam os pontos de desacordo e, caso não chegassem a um acordo, o terceiro revisor emitia uma decisão final baseada nos critérios de inclusão e exclusão predefinidos. Este processo assegurou a acurácia da seleção final dos artigos pelos revisores humanos.

Os critérios de inclusão adotados nesta revisão foram: (i) estudos que abordassem o desenvolvimento do Pensamento Computacional (PC) na educação em Química, ou que envolvessem formações de professores de Química com foco em PC; (ii) foco em públicos da área educacional, incluindo ensino básico, graduação, pós-graduação ou capacitações docentes; (iii) artigos publicados em inglês; e (iv) publicações com mais de 4 páginas, indexadas em anais de eventos ou periódicos nas bases de dados selecionadas.

É importante destacar que os dois revisores humanos não tiveram acesso aos resultados das IAs até a conclusão da seleção. Neste estudo, os resultados dos dois revisores humanos foram compilados em um resultado único, pois uma das premissas de uma RSL é contar com dois ou mais revisores independentes para minimizar os possíveis erros individuais.

As IAs, por outro lado, foram utilizadas de forma independente dos revisores humanos. Diferentemente dos humanos, essas ferramentas não sofrem fadiga cognitiva, mas é importante destacar que não foi conduzido um teste de reprodutibilidade formal para verificar se, ao submeter as mesmas entradas múltiplas vezes, os resultados permaneceriam exatamente os mesmos. Assim, embora presumivelmente consistentes, não é possível afirmar com total segurança que as IAs estão livres de variações no julgamento, especialmente considerando possíveis ajustes dinâmicos dos modelos baseados em contexto ou sessões. Outro ponto

importante é que embora os artigos selecionados sejam escritos na língua inglesa, essa não é a língua nativa dos revisores.

Neste estudo, as IAs não analisaram o texto completo dos artigos, pois o foco estava em avaliar as duas etapas iniciais de triagem, com o intuito de facilitar futuras revisões sistemáticas.

Na primeira etapa (leitura dos títulos), os artigos foram classificados conforme critérios previamente definidos, apresentados no Quadro 1.

Quadro 1 – Critérios de classificação na etapa de leitura dos títulos

Categoria	Descrição do critério
Artigo duplicado	Registros repetidos identificados em duas ou mais bases de dados.
Artigo secundário ou terciário	Estudos com delineamento de revisão sistemática, revisão bibliográfica ou taxonomia.
Resumo ou resumo expandido	Registros cujo título indicava tratar-se de resumo ou resumo expandido, sem análise do texto completo.
Não abordava STEM/STEAM, Química ou PC	Estudos que não contemplavam nenhum dos três eixos centrais da revisão.
Não é artigo	Registros cujo título indicava outro tipo de publicação (ex.: proceedings, poster, banner).
Dúvida	Registros cuja elegibilidade não pôde ser definida apenas pela leitura do título.
Aceito	Registros considerados potencialmente elegíveis após leitura do título.

Fonte: Elaborado pelos autores (2026)

Os artigos duplicados foram removidos previamente por um dos revisores antes do início do processo de triagem, permitindo que tanto os revisores humanos quanto as IAs iniciassem as leituras sem duplicações aparentes. No entanto, como algumas bases de dados aplicam variações de formatação ou remoção de caracteres especiais, duplicatas não identificadas inicialmente poderiam ainda estar presentes, razão pela qual o critério de identificação de artigos duplicados foi mantido na etapa de classificação pelos revisores.

Posteriormente, na segunda etapa os artigos aceitos ou em dúvida na primeira etapa eram classificados da seguinte forma:

- Não aceito: artigos rejeitados pelos revisores por não atenderem os critérios de inclusão.
- Dúvida: artigos em que os revisores ficaram em dúvida após a leitura do resumo.
- Aceito: artigos que os revisores consideraram potencialmente elegíveis.

Visando familiarizar as IAs com o tema da revisão sistemática, foi fornecido a elas um texto introdutório, usualmente referenciado como *prompt*, equivalente ao que os revisores humanos receberam, definindo a ação a ser executada. O prompt utilizado para instruir as ferramentas de IA nas etapas de triagem está apresentado no Quadro 2.

Quadro 2 – Prompt utilizado para a triagem com IA

Olá, A partir de agora você é um dos autores para construção de uma revisão sistemática. Para isso, você deverá passar por duas etapas: - Etapa 01 - Analisar os títulos dos artigos. Nesta etapa você receberá alguns títulos de artigos e deverá selecioná-los ou não para as próximas etapas, é importante que você não busque o artigo completo na internet e utilize apenas o título para seleção. - Etapa 02 - Analisar os resumos dos artigos selecionados na etapa 01. Também é importante que você leia apenas os resumos e não utilize fontes externas. Para a revisão sistemática foram utilizados artigos das seguintes bases de dados: IEEE, SpringerLink, ACM Library, ERIC, PubMed Wiley Online Library e EBSCO. A string de busca utilizada foi a seguinte: (“Computational thinking”) AND (”STEAM” OR “STEM”) AND (”Chemistry” OR “Chemical education”) Neste primeiro momento vamos realizar a etapa 01. Primeiramente você analisará apenas os títulos dos artigos, sem poder buscar na internet outras informações. A revisão sistemática selecionará artigos que utilizem Pensamento Computacional aplicado ao ensino de Química no contexto de STEAM, você poderá selecionar artigos que apliquem Química em qualquer nível de ensino (fundamental, médio ou graduação). Lembrando que alguns artigos podem utilizar ciência no local de Química. Para a primeira etapa, você deverá classificar os artigos em - Aceito - Dúvida - Artigo secundário ou terciário - aqui são artigos de revisão sistemática - Resumo expandido (menor ou igual a 4 páginas) - Artigo não utiliza STEAM ou Química ou Pensamento Computacional - Não é um artigo Lembrando: para os artigos serem selecionados os mesmos devem relacionar STEAM, Ensino de Química e Pensamento Computacional Para a segunda etapa os resumos serão divididos em aceitos, dúvida ou não aceitos. Podemos começar? Vou te enviar os títulos.

Fonte: Elaborado pelos autores (2026)

A base de dados de 1028 artigos únicos (após a remoção de duplicatas) foi formatada para a entrada nas IAs. Para o ChatGPT, os dados foram inseridos em grupos de até 50 títulos por vez, visando garantir um volume compatível com o desempenho da ferramenta e evitar limitações de processamento. Esse agrupamento é o que chamamos de blocos gerenciáveis, ou seja, conjuntos de dados que permitem à IA oferecer respostas coesas e completas. Inicialmente, apenas os títulos foram fornecidos; em seguida, os resumos foram apresentados da mesma forma. O mesmo procedimento foi adotado para o Gemini. A cada interação, a IA recebia o *prompt* com as instruções e um novo conjunto de títulos/resumos para avaliação, garantindo que a decisão fosse baseada exclusivamente nas informações fornecidas

Após a conclusão das duas etapas de triagem pelas IAs, os artigos selecionados por elas que não haviam sido incluídos pelos revisores humanos foram submetidos a uma terceira rodada de revisão por parte dos mesmos revisores humanos. Este processo de validação visou identificar se as IAs conseguiram capturar artigos relevantes que poderiam ter sido inadvertidamente perdidos na triagem inicial humana. Os revisores humanos, cegados para a origem da recomendação (ou seja, sem saber qual IA sugeriu o artigo), reavaliaram esses artigos utilizando os mesmos critérios de inclusão e exclusão definidos previamente. Este procedimento permitiu quantificar o valor complementar da IA na identificação de estudos

Os artigos foram, então, classificados da seguinte forma:

- Aceito: Artigo aceito pelas IAs, mas não selecionado pelos revisores humanos.
- Artigo não utiliza STEAM, PC e Química: Artigos que não relacionam as três áreas (STEAM, Pensamento Computacional e Química).
- Artigo teórico: Artigos secundários ou terciários, sem aplicação prática direta.
- Não é um artigo: Documentos como pôsteres, capas de anais de eventos e materiais semelhantes.

Além disso, para os artigos classificados como “Artigo não abordava STEAM, PC e Química”, os revisores avaliaram se o artigo abordava os temas STEAM, Pensamento Computacional, ou se pertencia a outra área de pesquisa.

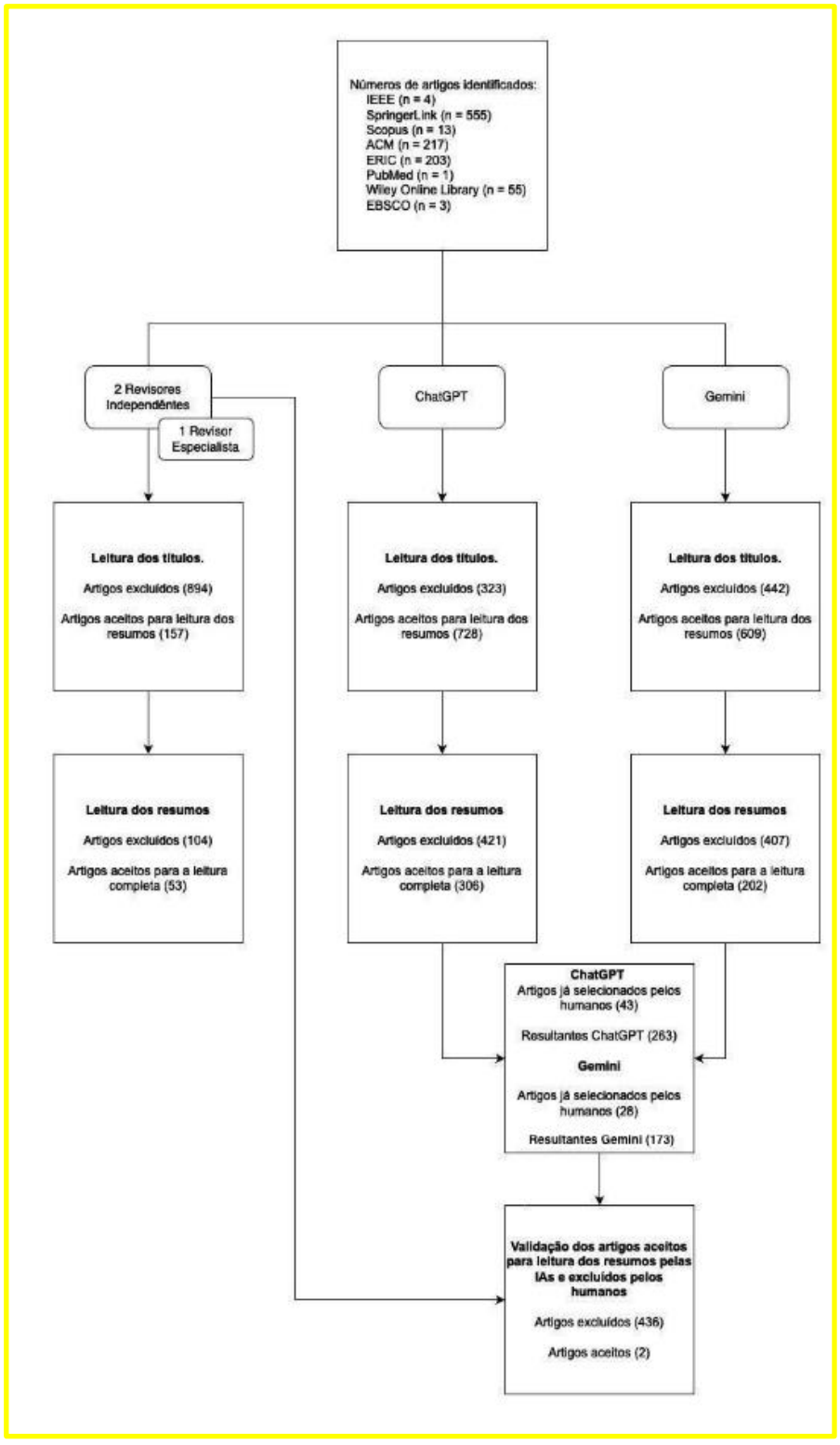
A análise dos dados foi de natureza comparativa e quantitativa. Os resultados de triagem dos revisores humanos e de cada IA foram compilados e comparados em termos de: (a) número total de artigos selecionados em cada etapa, (b) percentual de artigos incluídos/excluídos por categoria, e (c) interseção e exclusão entre as seleções (representadas por diagramas de *Venn*). A acurácia das IAs foi avaliada pela capacidade de identificar artigos previamente aceitos pelos humanos e, reciprocamente, pela identificação de artigos relevantes que os humanos não haviam selecionado. O tempo de execução das tarefas por humanos e IAs também foi comparado para demonstrar eficiência. As taxas de erro das IAs, especificamente a inclusão de artigos irrelevantes ou de outras categorias, foram quantificadas para analisar as limitações da automação.

Resultados e discussão

A Figura 1 mostra o fluxo de atividades realizadas neste estudo, detalhando as etapas conduzidas na RSL. A Figura 1 representa cada etapa executada pelos dois revisores independentes e pelas IAs operadas por um terceiro revisor.

A partir das buscas realizadas nas bases de dados selecionadas, foi obtido um total de 1051 artigos, importante destacar que os revisores humanos relavaram aproximadamente 4 meses para realizar todas as etapas, enquanto as IAs realizavam em um único dia. A Tabela 1 apresenta a distribuição do número de artigos retornados por cada base de dados, utilizando a string de busca escolhida.

Figura 1 – Fluxo das etapas realizadas.



Fonte: Elaborado pelos autores (2026)

Tabela 1 **–** Total de artigos selecionados em cada base.

BASE DE DADOS	RESULTADO
ACM	217
EBSCO	3
ERIC	203
IEEE	4
PUBMED	1
SCOPUS	13
SPRINGERLINK	555
WILEY ONLINE LIBRARY	55

Fonte: Elaborado pelos autores (2026)

A primeira etapa consistiu na eliminação dos artigos duplicados, resultando em 23 artigos que estavam disponíveis em mais de uma base de dados, restando assim 1028. Após essa etapa, os revisores e as IAs iniciaram as leituras dos títulos dos artigos. A Tabela 2 apresenta o total de títulos aceitos ou em dúvida, selecionados pelos revisores e pelos resultados das duas IAs.

Tabela 2 **–** Resultado da seleção da primeira etapa.

BASE DE DADOS	RESULTADO
REVISORES HUMANOS	157 (15%)
CHATGPT	728 (70%)
GEMINI	609 (59%)

Fonte: Elaborado pelos autores (2026)

Os revisores humanos eliminaram aproximadamente 85% dos artigos, nesta primeira etapa, enquanto as IAs selecionaram vários artigos que não continham termos relevantes para esta área: enquanto o ChatGPT refutou apenas 30% dos artigos, o Gemini refutou 41%. Dois possíveis cenários podem explicar este resultado: os humanos conseguiram eliminar vários estudos desnecessários já na primeira etapa; ou as IAs incluíram vários estudos que seriam importantes para a leitura.

Contudo, analisando esta primeira etapa, foi possível perceber que as IAs selecionaram títulos de periódicos, revistas e anais de eventos que os revisores humanos excluíram rapidamente. Isso ocorreu porque, mesmo quando o título continha expressões como “proceedings” ou “resumo expandido”, geralmente

indicativas de materiais fora dos critérios de inclusão, as IAs ainda assim consideraram esses registros como potencialmente relevantes, ao passo que os humanos os eliminaram prontamente com base nesses termos. Essa diferença evidencia uma limitação das IAs em aplicar automaticamente regras de elegibilidade com base em detalhes textuais mais específicos.

A Tabela 3 apresenta o total de artigos selecionados após a segunda etapa da RSL, a qual consistiu na realização da seleção de artigos pela leitura dos resumos dos artigos selecionados na primeira etapa. Nessa etapa, quando comparada a etapa inicial, os revisores humanos reduziram 67% dos artigos com base apenas nos títulos (restando apenas 5% dos artigos iniciais), enquanto o ChatGPT reduziu 68% (restando 30% dos artigos iniciais) e o Gemini, 67% (restando aproximadamente 20% dos artigos iniciais).

Tabela 3 **–** Resultado da seleção da segunda etapa.

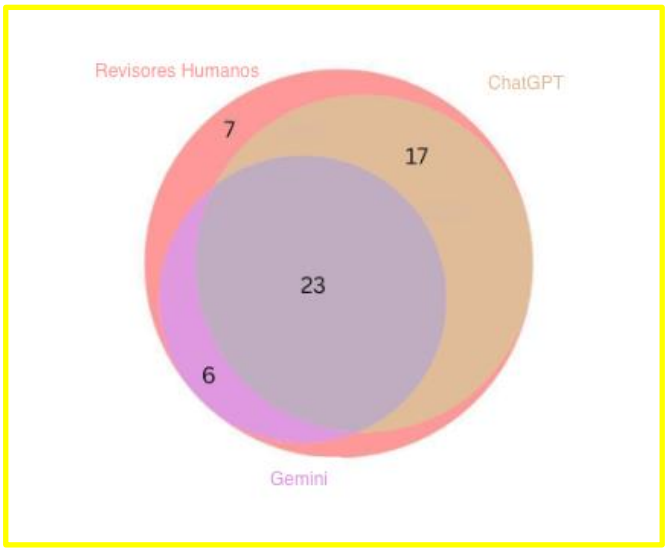
BASE DE DADOS	RESULTADO
REVISORES HUMANOS	53 ARTIGOS (Redução de 67% dos artigos comparado a etapa anterior)
CHATGPT	306 ARTIGOS (Redução de 68% dos artigos comparado a etapa anterior)
GEMINI	201 ARTIGOS (Redução de 67% dos artigos comparado a etapa anterior)
SEMELHANTES ENTRE AS IAS	125 ARTIGOS

Fonte: Elaborado pelos autores (2026)

A Figura 2 apresenta o diagrama de *Venn* com a comparação entre os 53 artigos selecionados pelos revisores humanos e os resultados obtidos pelas IAs. Observa-se que aproximadamente 43% dos artigos (23 artigos) foram selecionados tanto pelos humanos quanto pelas duas IAs. Ao comparar as IAs individualmente, ou seja, considerando apenas quando uma delas selecionou os artigos, o ChatGPT identificou cerca de 75% dos artigos escolhidos (40 artigos) pelos humanos, enquanto o Gemini selecionou apenas 54% (29 artigos). Sete artigos selecionados pelos humanos (aproximadamente 13%) não foram selecionados por nenhuma das IAs. Ao analisar os motivos, constatou-se que um desses artigos foi excluído já na primeira

etapa pelo ChatGPT. Os demais avançaram para a segunda etapa, sendo então excluídos por não apresentarem relação com o Pensamento Computacional aplicado à Química no resumo do trabalho. É importante destacar que, enquanto os humanos levaram cerca de 4 meses para concluir as duas primeiras etapas, as IAs realizaram o mesmo processo em apenas um dia, sendo que a maior parte do trabalho se deu na interação com as ferramentas do que no processamento propriamente dito.

Figura 2 **–** Diagrama de *Venn*: Artigos Selecionados pelos Revisores em comparação com as IAs.



Fonte: Elaborado pelos autores (2026)

Além da análise de sobreposição numérica apresentada anteriormente no diagrama de *Venn,* foi realizada uma comparação da representatividade relativa dessa interseção em relação ao total de artigos selecionados por cada IA após a segunda etapa. A Tabela 4 apresenta essa comparação entre o ChatGPT e o Gemini.

Tabela 4 **–** Comparativo entre os artigos selecionados pelas IA.

IA	TOTAL DE ARTIGOS	TOTAL DE ARTIGOS NA OUTRA IA
CHATGPT	306	121 (39%)
GEMINI	202	120 (59%)

Fonte: Elaborado pelos autores (2026)

Observa-se que, embora o ChatGPT tenha selecionado um número total maior de artigos, apenas 39% deles também foram selecionados pelo Gemini. Por outro lado, entre os artigos selecionados pelo Gemini, 59% também estavam presentes na

seleção do ChatGPT. Esses resultados indicam que a cobertura do Gemini foi proporcionalmente mais próxima da seleção do ChatGPT do que o contrário.

Na etapa final deste estudo, os revisores analisaram todos os artigos selecionados pelas IAs que não haviam sido previamente escolhidos por eles. Foram revisados 263 artigos selecionados pelo ChatGPT (306 menos os 43 que já haviam sido selecionados pelos humanos) e 173 artigos selecionados pelo Gemini (201 menos os 28 já selecionados pelos humanos). A Tabela 5 apresenta a classificação dos artigos selecionados pelas IAs e não pelos revisores. Os artigos são classificados como:

- Aceitos: artigos que anteriormente não foram aceitos pelos humanos e agora analisados novamente e aceitos para revisão sistemática. Importante ressaltar que pode ocorrer do mesmo artigo aceito ser contabilizado para ambas as IAs.
- Não relacionados a PC, STEAM e Química: quando o artigo não relaciona os três assuntos de forma simultânea.
- Artigo teórico: artigo secundário ou terciário.
- Não é um artigo: capa de anais de evento, pôster e resumos.

Tabela 5 **–** Comparativo entre os artigos selecionados pelas IA.

IA	TOTAL DE ARTIGOS	TOTAL DE ARTIGOS ANALISADOS	ACEITOS	NÃO RELACIONADOS A PC, STEAM E QUÍMICA	ARTIGO TEÓRICO	NÃO É UM ARTIGO
CHATGPT	306	263	2	238	53	24
GEMINI	202	173	1	142	12	18

Fonte: Elaborado pelos autores (2026)

Com a análise final dos artigos é possível perceber que os humanos deixaram de incluir 2 artigos para a seleção final da RSL (1 artigo comum entre as duas bases). Um dos artigos selecionados por ambas as IAs e não pelos humanos havia sido eliminado na primeira etapa por ambos os revisores humanos, o artigo de Bain, Dabholkar e Wilensky (2020) parece não apresentar no título a clareza do estudo “Confronting frame alignment in CT infused STEM classrooms” o que pode ter dificultado na seleção pelos revisores humanos. O outro artigo, selecionado apenas pelo ChatGPT, de Tarrés-Puertas *et al.* (2022) foi eliminado na segunda etapa pelos dois autores, ambos não identificaram a interdisciplinaridade entre Química e

Pensamento Computacional no artigo, assim, após a seleção das IA os revisores leram o artigo completo e identificaram a ligação entre as duas áreas, este segundo artigo deveria ter sido incluído como dúvida pelos autores para leitura completa do artigo, principalmente por trazer a multidisciplinaridade entre Química e Computação no resumo.

A não seleção de artigos essenciais para escrita da RSL é um cenário comum, como destacado por Wang *et al.* (2020), que analisaram revisões sistemáticas publicadas entre 2010 e 2017 e identificaram taxas significativas de erros em inclusões e exclusões realizadas por revisores humanos, com uma média de erro de 10%, variando entre 5% e 21%. Isso demonstra a vulnerabilidade dos revisores a tais equívocos. Além disso, Stoll *et al.* (2019) reforçam a importância de contar com pelo menos dois revisores independentes ao longo da condução de uma RSL. O estudo mostra que a presença de um segundo revisor em todas as etapas da triagem pode aumentar a identificação de estudos relevantes para a revisão sistemática.

A análise dos números permite concluir que ambas as IAs selecionaram muitos artigos que não relacionam as áreas de Pensamento Computacional e STEM/STEAM. No caso do ChatGPT, 37% dos artigos (99 artigos) não tinham nenhuma relação com PC e STEM/STEAM, enquanto no resultado do Gemini, 27% (48 artigos) não abordavam ambos os temas. A Tabela 6 apresenta esses dados de forma sumarizada.

Tabela 6 **–** Comparativo após avaliação entre os artigos selecionados pelas IA.

IA	TOTAL DE ARTIGOS	TOTAL DE ARTIGOS ANALISADOS	MENCIONAM PC	MENCIONAM STEM/ STEAM	MENCIONAM PC E STEM/ STEAM	NÃO MENCIONAM PC E STEM/ STEAM
CHATGPT	306	267	88	121	70	99
GEMINI	202	173	84	80	58	48

Fonte: Elaborado pelos autores (2026)

Também foi verificado se os artigos selecionados pelas IAs e não pelos revisores humanos tinham alguma relação com Pensamento Computacional ou STEAM. Dos 267 artigos analisados do ChatGPT, 121 artigos (aproximadamente 45%) tinham alguma relação com STEM/STEAM e 88 artigos (aproximadamente 32%) tinham relação com PC e 70 artigos (aproximadamente 26%) relacionavam PC e STEM/STEAM com outra área. Ao analisarmos os 173 artigos obtidos pelo Gemini, observamos que 80 artigos (aproximadamente 46%) tinham alguma relação com

STEM/STEAM, 84 artigos (aproximadamente 48%) estavam relacionados ao Pensamento Computacional (PC), e 58 artigos (aproximadamente 33%) abordavam ambos os temas.

Os resultados dessa experimentação mostram que ainda há uma grande discrepância entre os números obtidos pelos revisores humanos e pelas IAs, especialmente na etapa de leitura de resumos, quando tanto as IAs quanto os humanos têm acesso a mais conteúdo para a tomada de decisão. Em situações em que os humanos selecionaram 53 artigos, o ChatGPT selecionou 306 e o Gemini selecionou 201.

Com os resultados apresentados na Tabela 6, observou-se que as tecnologias utilizadas pelo ChatGPT e pelo Gemini influenciaram diretamente a triagem dos artigos nas duas primeiras etapas. A diferença na estrutura e nos objetivos dessas ferramentas pode explicar os resultados distintos observados na RSL. Enquanto o Gemini é focado em integração com ferramentas do ecossistema Google e se posiciona como uma IA versátil e informativa, o ChatGPT enfatiza a modelagem da linguagem natural e o refinamento na interpretação contextual (Rane; Choudhary; Rane, 2024).

Reforçando essa perspectiva, Issaiy et al. (2024) analisaram o desempenho do ChatGPT na triagem de resumos em revisões sistemáticas e identificaram que, embora a ferramenta apresente vantagens como economia de tempo e redução do esforço humano, ela ainda comete erros significativos, incluindo a seleção de artigos irrelevantes ou a exclusão de estudos relevantes. Esses achados destacam a importância da supervisão humana e da validação criteriosa ao empregar IAs na etapa de triagem, mesmo quando se trata de modelos linguísticos avançados.

Esses achados são corroborados por uma revisão sistemática recente Clark (2025), que avaliou o desempenho de diversas IAs generativas, incluindo ChatGPT, Copilot, DeepSeek, Dall-E e Gemini, nas etapas da síntese de evidências. Embora tenham demonstrado eficiência em tarefas específicas, como extração de dados em cenários simples, os resultados revelaram perdas significativas na recuperação de estudos relevantes (mediana de 91%) e taxas de erro consideráveis nas etapas de triagem (inclusão: até 29%; exclusão: até 83%). A conclusão dos autores reforça que, no estágio atual, essas ferramentas ainda não são confiáveis para conduzir revisões sistemáticas sem supervisão humana qualificada.

Além disso, é possível que ferramentas projetadas especificamente para a realização de revisões sistemáticas de literatura apresentem um desempenho superior. Uma destas ferramentas é o *ASReview*. Diferente do ChatGPT que utiliza o modelo GPT e do Gemini, que utiliza o modelo multimodal, o *ASReview* utiliza aprendizado ativo combinado com modelos de aprendizado supervisionado, como Random Forest, Regressão Logística e Naive Bayes, para priorizar artigos relevantes com base em feedback iterativo do usuário durante a triagem de estudos. No estudo dos autores Quan, Tytko e Hui (2024), os autores avaliaram a utilização desta ferramenta em um ambiente controlado. Os resultados indicam que seu uso pode facilitar a triagem; contudo, os autores alertam que, mesmo com o uso de uma ferramenta, ainda existem limitações. É imprescindível que os dados sejam corretamente configurados na entrada da ferramenta, e o resultado dependerá diretamente da experiência do pesquisador.

Outro ponto a ser observado é a importância de indicar informações relevantes ao estudo relatado no artigo em seu título, resumo e mesmo nas palavras chaves associadas ao trabalho, conforme recomendado pelo IMRaD (Shiely; Gallafher; Millar 2024) e mesmo observado nos trabalhos que orientam a execução de RSLs (Sollaci; Pereira, 2004). Assim, tanto os revisores humanos quanto as IAs conseguem identificar de forma mais clara quais estudos devem ou não ser selecionados para as etapas seguintes. Pottier *et al.* (2024) avaliaram diversos estudos e identificaram que em vários cenários os artigos não foram encontrados em buscas dos pesquisadores por não utilizarem diretrizes e técnicas que possam auxiliar na descoberta do artigo científico.

Também é perceptível que o cansaço dos revisores e o "olho treinado" podem levar à desclassificação de estudos durante as etapas iniciais, quando o volume de trabalhos é muito grande, além da extração de dados nos artigos, inclusive em revisores mais experientes (Jyu *et al*., 2020).

Portanto, é altamente recomendado que mais estudos utilizando IA nas etapas de triagem de uma RSL sejam explorados, com o objetivo de encontrar soluções para as fases iniciais, que são cansativas e demandam muito tempo dos revisores. A reprodutibilidade das soluções de IA utilizando as mesmas bases de dados também deve ser um ponto central a ser avaliado, pois ela garante que os resultados obtidos em uma RSL possam ser replicados e validados em diferentes contextos e por

diferentes equipes. Isso é essencial para a confiabilidade do uso de tecnologias e métodos no meio acadêmico, especialmente quando aplicadas em processos como recuperação de conteúdos e triagem (Dennstadt, 2024). Recomenda-se, também, que um número crescente de IAs seja testado e avaliado no meio acadêmico, a fim de que essas tecnologias se tornem, cada vez mais, ferramentas úteis para os pesquisadores em etapas como a recuperação de conteúdos, redação e compreensão de texto (Shukla *et al.,* 2024).

Limitações e direções futuras

Ao longo desta pesquisa, ficou evidente que as IAs desempenham um papel extremamente eficiente nas fases introdutórias de uma RSL, principalmente ao otimizar o tempo gasto nas etapas iniciais. Entretanto, mesmo com o uso de uma IA para facilitar a triagem dos estudos, é fundamental manter a supervisão de um revisor humano e garantir que os dados inseridos na IA sejam de qualidade.

Neste estudo, avaliamos duas IAs com tecnologias diferentes. Recomenda-se que estudos futuros testem outras IAs nas etapas de triagem de uma RSL, como o *Microsoft Copilot*, que utiliza a mesma tecnologia empregada no ChatGPT. Assim, será possível avaliar se uma determinada tecnologia apresenta um desempenho superior às demais. Outra possibilidade de investigação futura relevante é a comparação entre o desempenho de revisores humanos, IAs específicas para revisões sistemáticas (*DistillerSR* e *Rayyan*) e IAs de uso geral (como ChatGPT, Copilot e Gemini).

Conclusão

Ao longo desta pesquisa, verificou-se que as IAs reduziram significativamente o tempo de triagem inicial, realizando em um dia uma tarefa que demandou aproximadamente quatro meses de trabalho humano. Esse resultado evidencia seu potencial para acelerar etapas preliminares de Revisões Sistemáticas de Literatura.

Entretanto, foram observadas limitações relevantes, especialmente a elevada taxa de inclusão de documentos irrelevantes e a seleção de itens que não configuravam artigos científicos. Esses achados indicam que o uso de IAs na triagem ainda exige supervisão humana para garantir rigor metodológico. Por outro lado, as

ferramentas identificaram dois artigos elegíveis não selecionados pelos revisores, demonstrando potencial complementar ao julgamento humano.

Conclui-se que as IAs podem atuar como apoio estratégico nas fases iniciais de RSLs, desde que utilizadas como filtro preliminar e acompanhadas de validação criteriosa. Entre as limitações do estudo, destacam-se o uso de versões gratuitas das ferramentas, a ausência de análises estatísticas aprofundadas e a aplicação em um único recorte temático. Pesquisas futuras poderão ampliar a comparação entre modelos e contextos disciplinares, contribuindo para o uso mais preciso e ético da IA em revisões científicas.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 00

Referências

AROMATARIS, E.; PEARSON, A. The systematic review: an overview. **AJN The American Journal of Nursing**, v. 114, n. 3, p. 53-58, 2014.

BAIN, C.; DABHOLKAR, S.; WILENSKY, U. Confronting frame alignment in CT infused STEM classrooms. **International Conference on Computational Thinking Education** v. 91: p. 91–94, 2020.

BLAIZOT, A. *et al.* Using artificial intelligence methods for systematic review in health sciences: A systematic review. **Research Synthesis Methods**, v. 13, n. 3, p. 353-362, 2022.

BORAH, R. *et al.* Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. **BMJ open**, v. 7, n. 2, p. e012545, 2017.

CLARK, J. *et al.* Generative artificial intelligence use in evidence synthesis: A systematic review. **Research Synthesis Methods**, p. 1-19, 2025.

DENNSTÄDT, F. *et al.* Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. **Systematic Reviews**, v. 13, n. 1, p. 158, 2024.

SANTOS, J. J. dos; SANTO, E. do E. Produto educacional como proposta formativa para a integração das tecnologias digitais no contexto da cultura digital. **Educitec Revista de Estudos e Pesquisas sobre Ensino Tecnológico**, Manaus, Brasil, v. 11, n. jan./dez., p. e256825, 2025.

EVANGELISTA, A. H. A. *et al.* Impactos da incorporação da Inteligência Artificial no ensino de Matemática: um Estado do Conhecimento. **Educitec Revista de Estudos e Pesquisas sobre Ensino Tecnológico**, v. 11, n. jan./dez., p. e265425-e265425, 2025.

JYU, J. *et al.* Adjudication rather than experience of data abstraction matters more in reducing errors in abstracting data in systematic reviews. **Research synthesis methods**, v. 11, n. 3, p. 354-362, 2020.

KACENA, M. A.; PLOTKIN, L. I.; FEHRENBACHER, J. C. The use of artificial intelligence in writing scientific review articles. **Current Osteoporosis Reports**, v. 22, n. 1, p. 115-121, 2024.

KHALIFA, M.; ALBADAWY, M. Using artificial intelligence in academic writing and research: An essential productivity tool. **Computer Methods and Programs in Biomedicine Update**, v. 5, p. 100 145, 2024.

MENTA, E.; BRITO, G. S. O Papel da Inteligência Artificial no Ensino Tecnológico: Implicações Emergentes. **Educitec Revista de Estudos e Pesquisas sobre Ensino Tecnológico**, Manaus (AM), v. 10, e232524, 2024.

PAGE, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **BMJ**, v. 372, n. 71, p. 1-9. https://doi.org/10.1136/bmj.n71. 2021.

POTTIER, P. *et al.* Title, abstract and keywords: a practical guide to maximize the visibility and impact of academic papers. **Proceedings B**, v. 291, n. 2027, p. 20241222, 2024.

QUAN, Y.; TYTKO, T.; HUI, B. Utilizing AS Review in screening primary studies for meta-research in SLA: A step-by-step tutorial. **Research Methods in Applied Linguistics**, v. 3, n. 1, p. 100-101, 2024.

SHIELY, F.; GALLAGHER, K.; MILLAR, S. R. How, and why, science and health researchers read scientific (IMRAD) papers. **Plos one**, v. 19, n. 1, p. e0297034, 2024.

SHUKLA, M. *et al.* A comparative study of ChatGPT, Gemini, and Perplexity. **International Journal of Innovative Research in Computer Science & Technology**, v. 12, n. 4, p. 10-15, 2024.

SOLLACI, L. B.; PEREIRA, M. G. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. **Journal of the medical library association**, v. 92, n. 3, p. 364, 2004.

STOLL, C. RT *et al.* The value of a second reviewer for study selection in systematic reviews. **Research synthesis methods**, v. 10, n. 4, p. 539-545, 2019.

RANE, N.; CHOUDHARY, S.; RANE, J. Gemini versus ChatGPT: applications, performance, architecture, capabilities, and implementation. **Performance, Architecture, Capabilities, and Implementation** , v.5, n 1, p. 69-93, 2024.

TARRÉS-PUERTAS, M. I. *et al.* Sparking the interest of girls in computer science via chemical experimentation and robotics: The Qui-Bot H2O case study. **Sensors**, v. 22, n. 10, p. 3719, 2022.

TRAD, F. *et al.* Streamlining systematic reviews with large language models using prompt engineering and retrieval augmented generation. **BMC medical research methodology**, v. 25, n. 1, p. 130, 2025.

WANG, Z. *et al.* Error rates of human reviewers during abstract screening in systematic reviews. **PloS one**, v. 15, n. 1, p. e0227742, 2020.

WILLIAM, F. K. A. AI in academic writing: Ally or foe. **International Journal of Research Publications**, v. 148, n. 1, 2024.

VAN DIJK, S. H. *et al.* Artificial intelligence in systematic reviews: promising when appropriately used. **BMJ open**, v. 13, n. 7, p. e072254, 2023.

Recebido: 23/06/2025
Aprovado: 18/02/2026
Publicado: 26/02/2026
Como citar (ABNT): PUNTEL, F. E. *et al*. Humanos vs. inteligências artificiais em revisões sistemáticas: avaliação nas etapas de seleção por títulos e resumos. **Educitec - Revista de Estudos e Pesquisas sobre Ensino Tecnológico,** Manaus, v. 12, e270726, 2026
Contribuição de autoria:
Fernando Emilio Puntel: Conceituação; Curadoria de Dados; Investigação; Metodologia; Administração de Projeto; Supervisão; Validação; Visualização; Escrita (rascunho original).
Muriel Belo Pereira: Conceituação; Curadoria de Dados; Investigação; Escrita (rascunho original); Escrita (revisão e edição).
Bruna Adriane Fary: Conceituação; Investigação; Escrita (revisão e edição).
Gerson Geraldo Homrich Cavalheiro: Conceituação; Investigação; Metodologia; Validação; Escrita (revisão e edição)
Editor responsável: Iandra Maria Weirich da Silva Coelho